COLLABORATION BETWEEN UK UNIVERSITIES: A MACHINE-LEARNING BASED WEBOMETRIC ANALYSIS

Patrick Tarilayefa Kenekayoro MSc

A thesis submitted in partial fulfilment of the

requirements of the University of Wolverhampton

for the degree of Doctor of Philosophy

September 2014

Signature……..

Date……..

Abstract

Collaboration is essential for some types of research, which is why some agencies include collaboration among the requirements for funding research projects. Studying collaborative relationships is important because analyses of collaboration networks can give insights into knowledge based innovation systems, the roles that different organisations play in a research field and the relationships between scientific disciplines.

Co-authored publication data is widely used to investigate collaboration between organisations, but this data is not free and thus may not be accessible for some researchers. Hyperlinks have some similarities with citations, so hyperlink data may be used as an indicator to estimate the extent of collaboration between academic institutions and may be able to show types of relationships that are not present in co-authorship data. However, it has been shown that using raw hyperlink counts for webometric research can sometimes produce unreliable results, so researchers have attempted to find alternate counting methods and have tried to identify the reasons why hyperlinks may have been created in academic websites.

This thesis uses machine learning techniques, an approach that has not previously been widely used in webometric research, to automatically classify hyperlinks and text in university websites in an attempt to filter out irrelevant hyperlinks when investigating collaboration between academic institutions.

Supervised machine learning methods were used to automatically classify the web page types that can be found in Higher Education Institutions' websites. The results were assessed to see whether

automatically filtered hyperlink data gave better results than raw hyperlink data in terms of identifying patterns of collaboration between UK universities.

Unsupervised learning methods were used to automatically identify groups of university departments that are collaborating or that may benefit from collaborating together, based on their co-appearance in research clusters.

Results show that the machine learning methods used in this thesis can automatically identify both the source and target web page categories of hyperlinks in university websites with up to 78% accuracy; which means that it can increase the possibility for more effective hyperlink classification or for identifying the reasons why hyperlinks may have been created in university websites, if those reasons can be inferred from the relationship between the source and target page types.

When machine learning techniques were used to filter hyperlinks that may not have been created because of collaboration from the hyperlink data, there was an increased correlation between hyperlink data and other collaboration indicators. This emphasises the possibility for using machine learning methods to make hyperlink data a more reliable data source for webometric research.

The reasons for university name mentions in the different web page types found in an academic institution's website are broadly the same as the reasons for link creation, this means that classification based on inter-page relationships may also be used to improve name mentions data for webometrics research.

Clustering research groups based on the text in their homepages may be useful for identifying those research groups or departments with similar research interests which may be valuable for policy makers in monitoring research fields; based on the sizes of identified clusters and for identifying future collaborators; based on co-appearances in clusters, if identical research interests is a factor that can influence the choice of a future collaborator.

In conclusion, this thesis shows that machine learning techniques can be used to significantly improve the quality of hyperlink data for webometrics research, and can also be used to analyse other web based data to give additional insights that may be beneficial for webometrics studies.

Table of Contents

Acknowledgements

I would like to thank my supervisors Dr. Kevan Buckley and Prof. Mike Thelwall for their advice and guidance on my research throughout my studies. They made the process significantly easier than it could have been.

I would like to thank all the members of the Statistical Cybermetrics Research Group for their helpful feedback, especially during the doctoral forums.

I'd like to thank Ludo Waltman, for providing the co-authored publication data for UK universities.

I also would like to thank my family and friends, especially my parents Mr and Mrs Kenekayoro for their support and prayers throughout my studies.

Publications from thesis

Journal Papers:

Kenekayoro, P., Buckley, K. and Thelwall, M. (2014) Automatic classification of academic web page types, *Scientometrics*, Springer Netherlands, pp. 1–12, [online] Available from: http://dx.doi.org/10.1007/s11192-014-1292-9.

Kenekayoro, P., Buckley, K. and Thelwall, M. (2014b) Hyperlinks as inter-university collaboration indicators, *Journal of Information Science*, **40**(4), pp. 514–522, [online] Available from: http://jis.sagepub.com/content/40/4/514 (Accessed 18 July 2014).

Kenekayoro, P., Buckley, K. and Thelwall, M. (In Press) Clustering Research Group Website Homepages, *Scientometrics*, Springer Netherlands.

Conference Papers:

Kenekayoro, P., Buckley, K. and Thelwall, M. (2012) Fuzzy Clustering of UK Computer Science Departments, In IADIS European Conference on Data Mining (DM), Ries, A. P. dos, Wang, P. S. P., and Abraham, A. P. (eds.), Lisbon, pp. 203–208.

Kenekayoro, P., Buckley, K. and Thelwall, M. (2013) Motivation for Hyperlink Creation Using Inter-Page Relationships, In 14th Conference of the International Society for Scientometrics and.

Informetrics (ISSI), Gorraiz, J., Schiebel, E., Gumpenberger, C., Hörlesberger, M., and Moed, H. (eds.), Vienna, pp. 1253–1269.

List of Figures

List of Tables

1. Introduction

Webometrics has been defined as using different methods to study web content for social science goals (Thelwall, 2009). The web is the largest freely available data source, so analysing this data opens the possibility for identifying previously unknown trends and can be an alternative for other data sources that may be unavailable or expensive to access for some researchers. Although any type of web data can be used in webometrics research, for example, hyperlinks or text have been used in a number of studies, link based data was widely used in early webometrics studies because links have some similarities with citations in academic articles (Rousseau, 1997). Early webometrics research used bibliometric methods for their web based studies.

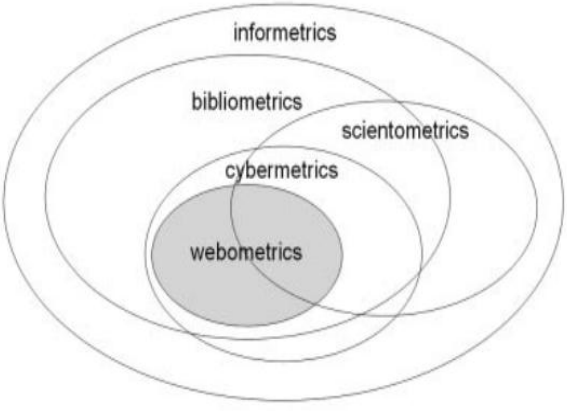

Figure 1.1 Relationships between informetric fields (Björneborn and Ingwersen, 2004)

Björneborn and Ingwersen (2004) described the relationship between different informetric fields. Informetrics is the study of any

kind of information, not just publications and not restricted to science while bibliometrics is the qualitative analysis of publications (Hérubel, 1999) and measures and analyses publication patterns of scientific texts, books or information in different disciplines. Scientometrics not only analyses the bibliographic data of scientific publications but also analyses all scientific activities, from production to dissemination and use (Jacobs, 2010). Cybermetrics aims to measure the internet, while webometrics measures the web (rather than the whole internet) using bibliometric and informetric techniques.

A number of webometric studies (Holmberg and Thelwall, 2009; Minguillo and Thelwall, 2012; Boell, Wilson and Cole, 2008; Vaughan, Tang and Du, 2009; Minguillo and Thelwall, 2011; Ortega et al., 2008; Park and Thelwall, 2008; Thelwall and Zuccala, 2008; Vaughan, Kipp and Gao, 2007; Bar-Ilan, 2004) have used hyperlinks as the main data source for their research. They have been used to study academic websites, for example to show that there is a relationship between the distance separating two universities and the number of links between these two universities (Thelwall, 2002d). The in-link count to a UK university website also associates with the research assessment rating of that university (Xuemei et al., 2003), which is also in line with the concept of hubs and authorities, where highly linked websites are seen as authorities in a particular domain.

Link relationships between websites have been used in information retrieval systems. Two popular algorithms, Hyperlink Induced Topic Search algorithm (HITS) (Kleinberg, 1999) and Google's PageRank (Page et al., 1999) use the similarity between hyperlinks and citations in academic publications to rank web pages in response to a search.

Often using hyperlink-based ranking algorithms themselves, web search engines now serve as a key source of the data used in webometric studies. Search engines crawl the web, store web pages in a database and retrieve relevant results for users depending on their search queries. Although search engines were not designed with webometrics research in mind, they have particular features that aid webometric studies or in some cases make webometric studies possible. It is not practical for a single researcher to design a personal crawler that retrieves data from the whole web, so search engines are used as tools for data collection in the webometric researches that need to get data from the whole web, or as much of it as possible. For example, co-link data for a group of organisations' websites from search engines can be used for business intelligence (Vaughan and You, 2008) and search engines can also be used to identify emergent trends or to track trends, based on its first appearance on the internet (Thelwall and Hasler, 2007; Chen, Tsai and Chan, 2007).

The majority of webometric studies collect data using search engines or a personal web crawler and then analyse this data using a

suitable method for the research goals. This thesis primarily uses machine learning techniques in conjunction with more standard methods for data analysis to reach the goal of using web-based data to investigate collaborative relationships between academic institutions. Machine learning techniques identify patterns in a data set using complex heuristics or mathematical functions. They have been applied to a number of research fields, including biology, natural language processing (NLP), and economics, but are not yet widely adopted in the webometrics research community. In this thesis, machine learning techniques are used to make hyperlink analysis more effective for large scale webometrics studies and to automatically group similar organisations that may benefit from collaborations.

1.1. Aim

Broadly, the goal of this thesis is to demonstrate the value of machine learning techniques for webometrics research, as few webometrics studies have used this method. Four studies aimed to address some webometrics research issues are used to achieve this broad goal. These aims are (a) identify an effective way to infer the reasons for link creation in academic websites, which in turn will make large scale hyperlink analyses more effective, (b) improve the quality of web based data to make it more suitable for studying collaborative ties among UK

academic organisations and (c) investigate what could be achieved through clustering the text in academic web pages; the majority of early webometric analysis of academic websites have focused on hyperlink relationships.

The subsequent studies have specific research questions that help reach the aims of the thesis.

1.1.1. Study One: Automatically classifying web pages and hyperlink targets

Despite the extensive use of hyperlink data in webometric research, hyperlink data can be very unreliable (Thelwall, 2002e) and it is difficult to identify why hyperlinks have been created in academic websites. For this reason, alternate methods for counting links instead of using raw link counts have been investigated (Thelwall and Wilkinson, 2008) and several other studies have attempted to find the reasons for link creation in Higher Education Institutions' (HEIs) websites (Bar-Ilan, 2005; Wilkinson et al., 2003; Stuart, Thelwall and Harries, 2007). These studies manually classify individual links, but this approach is infeasible for large scale studies because it takes too long. A more effective approach is needed and Stuart, Thelwall and Harries (2007)

suggest that methods for automatic classification of links should be developed for the potential of links to be fully harnessed in webometrics research.

The first study in this thesis attempts to address this by answering the following research questions:

- How reliably can machine learning techniques automatically classify university web page types?
- How reliably can machine learning techniques predict the classification of link target pages from characteristics of link source pages?
- What are the common characteristics of out-links from each university web page type?

The reasons why a link may have been created in a website can be inferred by studying the two pages that the hyperlink connects. If university web pages are grouped into categories, the relationship between the aggregate of the links connecting two categories may also reflect that of the links from individual web pages.

Answering the research questions in this study will show if the reasons for link creation can be automatically identified using machine learning methods. If this is possible, then the links that may not have been created because of collaboration reasons can be automatically excluded, which in turn may improve the quality of hyperlink data for

studying collaboration. The next study investigates the extent to which excluding certain links may be able to improve hyperlink data for collaboration studies.

1.1.2. Study Two: Using machine learning to filter out irrelevant links for collaboration studies

Collaboration may be beneficial for research, which is why some funding agencies include collaboration as one of the requirements for funding research projects (Lee and Bozeman, 2005; Sonnenwald, 2007). Even though collaborations may not always lead to a published article and researchers collaborating together may not necessarily co-author an article, co-authored publications are widely used to study collaboration between researchers. At a higher level, institutional collaboration can also be investigated through the researchers' organisational affiliations in published articles.

Traditional collaboration studies use co-authorship information extracted from publication metadata retrieved from databases like Thomson Reuters Web of Knowledge and Scopus. This data can sometimes be expensive to access, which makes it unavailable for researchers without the necessary funds. Web hyperlinks are an

alternative data source. Even though hyperlink counts may be less accurate than ISI (Institute of Scientific Information) database data because the majority of hyperlinks in academic websites are not created because of collaboration (Stuart, Thelwall and Harries, 2007), they can be particularly useful for pilot studies that can indicate if it is worth investing time and money for a full scale investigation.

The second study in this thesis investigates whether machine learning techniques to filter out irrelevant hyperlinks can be used to investigate the extent to which two universities collaborate together. The following research questions are answered to reach this goal:

- Can the extent of collaboration between two universities be better estimated with hyperlinks if only those links between university staff web pages are used rather than all links?
- Can the extent to which a university collaborates with other UK universities be better estimated by the total number of academic in-links rather than the total in-links to the university's website?

Machine learning techniques are used to automatically identify academic in-links and links from university staff web pages.

Addressing the research questions in this study will show if subsets of links (academic in-links and links from university staff pages) are more suitable for studying collaboration between academic

institutions rather that all links, and if machine learning methods can effectively identify these subsets of links.

1.1.3. Study Three: Investigating web mentions as a collaboration indicator

Hyperlink data was predominantly used in early webometrics research because of its conceptual albeit superficial similarity with citations in bibliometrics (Ingwersen and Björneborn, 2005). Hyperlinks are still different from citations. One of the first definitions of webometrics was applying bibliometric and informetric techniques to web data (Almind and Ingwersen, 1997). Now, co-word occurrences (Vaughan and You, 2010), URL citations (Kousha and Thelwall, 2007) and web mentions or organisation name mentions in web pages (Cronin et al., 1998) have been introduced in different webometrics studies.

Thelwall and Sud (2011) showed that there is significant correlation between the number of web mentions of an organisation and in-link counts to that organisation's website, so web mentions can also be used for organisation impact studies, but do web mentions perform as well or better than URL counts for collaboration studies?

The third study in this thesis investigates if the web mentions in universities' websites can be used as an alternative to hyperlinks to study the extent to which two universities collaborate together by answering the following research questions:

- Does the number of mentions of a university's name in another university's website correlate with the traditional indicators of collaboration between the two universities?
- Is the correlation between university name mentions and the traditional indicators of collaboration higher than the correlation between hyperlink counts and the traditional indicators of collaboration?
- What are the reasons for university name mentions in the different web page categories that could be found in a typical university's website?

In addressing the research questions in this study, the suitability of using university name mentions to study collaboration, and if the machine learning techniques that can be applied to hyperlink data can be transferred to name mentions data will be identified. This is particularly useful because if name mentions are as reliable as hyperlink data, some ethical concerns regarding hyperlink data collection through web crawling may be avoided.

1.1.4. Study Four: Exploratory cluster analysis of computer science research groups

Previous webometrics studies have conducted exploratory cluster analyses with software like Pajek (de Nooy, Mrvar and Batagelj, 2005) that exploit the graph structure of hyperlink relationships between websites (Ortega and Aguillo, 2009; Thelwall, 2001a; Meloche, 2010; Onyancha and Ocholla, 2007; Thelwall and Zuccala, 2008; Minguillo and Thelwall, 2011). François, Lamirel and Shehabi (2008) argue that graph theoretic methods cannot reliably support accurate analysis based on multiple factors, and proposed an alternate clustering solution based on Self-Organising Maps (SOM). SOM is one of several unsupervised machine learning techniques.

The fourth study in this thesis describes methods that use unsupervised machine learning algorithms for exploratory webometrics research, driven by the following research question:

- Can an unsupervised machine learning cluster analysis of the text in the homepages of computer science research groups in the UK with self-organising maps and principal component analysis reflect similarities in interests between the departments?

In addressing the research question in this study, the relevance of clustering the text in the homepages of university websites will be identified. This is useful because the majority of academic webpage cluster analysis have focused on hyperlink relationships. This study will give insights to what can be identified through co-word analysis.

Text analyses can be useful to give context to webometric or bibliometric studies. Co-words gave similar results to citations when Leydesdorff (1989) analysed a set of biochemistry articles with factor analysis and hierarchical cluster analysis. In bibliometrics, it is widely used to map scientific fields (Heimeriks and van den Besselaar, 2006; Peters and van Raan, 1993; Janssens et al., 2006; Whittaker et al., 1989). Co-word analyses can show mappings of research topics in terms of the concepts used, with their contextual meanings identified through the associated cited references (van den Besselaar and Heimeriks, 2006).

Co-word analysis has not been used extensively to analyse academic websites, although it has been successfully applied in the study of triple helix relationships (university – industry – government) on the web between organisations (Khan and Park, 2011).

1.2. Thesis Outline

This thesis is divided into seven chapters. After the introductory chapter, Chapter Two gives a detailed review of important webometric studies and a description of several machine learning techniques. The similarities between web mining and webometrics is discussed. The application areas of link analysis and where webometric studies are used to investigate relationships between organisations are reviewed, and then the machine learning algorithms used in this thesis are described.

Chapter Three discusses methods used in webometric studies and the techniques used in this thesis. A section in chapter three describes methods for data collection for webometrics studies and when it is suitable to utilise search engines or design a personal web crawler for data collection. Webometrics research is primarily quantitative, so the statistical techniques used in this thesis to find the relationship between quantitative variables and to visualize clusters of identical elements are discussed. Procedures for successful supervised machine learning are described along with accuracy measures that determine if a classification model can correctly predict future unseen cases after training with an initial set (training set).

Chapter Four describes the methods and results of the first study, concerned with "identifying an effective method for hyperlink classification on a large scale". Categories that university web pages can belong to are identified and then supervised machine learning techniques are used to assign web pages into categories. Supervised learning is also used to predict the target of a hyperlink from information available in the source page. The reasons for out-links in different web page types are also investigated. Part of this chapter has been published in a journal (Kenekayoro, Buckley and Thelwall, 2014a) and presented in a conference (Kenekayoro, Buckley and Thelwall, 2013).

Chapter Five describes the methods and results of the second and third studies. Hyperlink data and web mentions are compared with data used in traditional collaborative studies to determine if there is any correlation between these data sets. A supervised machine learning technique is used to filter out hyperlinks from unwanted web page types in order to identify if restricting links to only those hyperlinks coming from particular web page types increases the correlation between hyperlinks and other collaboration indicators. Random selections of links are studied to identify reasons for web mentions in different web page types. Part of this chapter has been published in a journal (Kenekayoro, Buckley and Thelwall, 2014b).

Chapter Six describes the methods and results of the fourth study. Clustering techniques are used to analyse web data for webometrics research. The text in homepages of computer science research groups are used for cluster analysis in order to identify key research areas in computing and to find similar departments based on the co-word occurrences of computer science keywords in the homepages of their research groups. Part of this chapter was presented in a conference (Kenekayoro, Buckley and Thelwall, 2012) and has been accepted for publication in Scientometrics' journal.

Chapter Seven summarises the thesis, describing the main goals, and how they were achieved through the case studies. Key findings from each study in the thesis are summarised and the last sections state what this thesis contributes to webometrics research, summarises its limitations, and proposes further work.

2. Literature Review

This chapter starts off with a description of research collaboration and reviews the traditional ways (with bibliometric data) in which collaboration among organisations is investigated. The similarity between hyperlink data and bibliometric data is discussed, as this is what motivated webometrics research field. Section 2.2 describes web mining, and states the similarity of this computer science research field with webometrics, particularly in the tasks carried out in both webometrics and web mining.

Section 2.3 reviews application areas of webometrics research, with particular emphasis on academic web analyses as this thesis is about studying collaboration between academic organisations.

Section 2.4 gave a description of the supervised and unsupervised machine learning techniques that were used in subsequent empirical chapters. Traditional clustering techniques (partitional and hierarchical) as well as graph clustering was briefly discussed because they are widely used in webometrics research. This section did not go in depth into all machine learning algorithms. It only gives an overview of the working principles of the algorithms.

Section 2.5 summarises the key points from all sections in the literature review.

2.1. Introduction

Research collaboration is defined as *"the coming together of researchers to achieve the common goal of producing new scientific knowledge"* (Katz and Martin, 1997). In the worst case collaboration does not positively influence research productivity and in the best case it increases it (Beaver, 2001). However, the main goal of research collaboration is the creation of new scientific knowledge (Katz and Martin, 1997) and collaborative academic papers have been shown to have more impact than single authored papers (Katz and Hicks, 1997; Amin and Mabe, 2000). Studies have shown collaboration to occur for many reasons, some of which are access to expertise, equipment or funds (Beaver, 2001). Collaboration has also been shown to improve research productivity in terms of the number of publications produced (Lee and Bozeman, 2005; Landry, Traore and Godin, 1996; Katz and Martin, 1997; Subramanyam, 1983).

There is a consensus about the importance of scientific collaboration for researchers. Some government funding agencies encourage research collaboration by adding collaboration to part of the funding requirements (Lee and Bozeman, 2005; Sonnenwald, 2007)

and collaboration associates with research productivity in terms of the number of publications (Lee and Bozeman, 2005; Landry, Traore and Godin, 1996; Katz and Martin, 1997; Subramanyam, 1983). Even though inter-organisational collaboration is encouraged, the majority of research projects have only one participant. In 2012, approximately 60% of EPSRC funded projects awarded to UK universities in chapter 5 had only one participating organisation. This is in line with other results that the majority of collaboration is within a single organisation (Gazni, Sugimoto and Didegah, 2012).

Studying collaboration is important for exploring the relationships between organisations, which can aid in identifying important or influential actors and the role that different organisations play in a particular research field. Collaboration studies can also be used to explore knowledge based innovation systems (Stuart, 2008). Knowledge based innovation systems are *"systems where efficient interactions between actors enable greater innovation"* (Stuart, 2008; Potratz and Widmaier, 1996).

Researchers have analysed the collaboration networks of organisations that participated in EU funded projects using statistical and social network analysis techniques (Ortega and Aguillo, 2010b; Roediger-Schluga and Barber, 2008; Ortega and Aguillo, 2010a). Results from these studies have given an overview of the main properties of the collaboration network of different research fields and

the roles organisations play in the knowledge innovation system. For example, the core of the network in technical fields has a high proportion of large companies because organisations in these sectors are interested in projects with high profit potential, while the core of the network in health related fields is made up of universities and research centres because of the social importance of health (Ortega and Aguillo, 2010b) or perhaps, pharmaceutical companies are secretive because of the vast amount of money to be made, and hence do not publish much.

A number of studies have investigated collaboration between individuals, departments or organisations through co-authored publications. Based on co-authorship, European researchers collaborate more with global researchers than with exclusively European researchers (Mattsson et al., 2010), although collaboration within a single institution still produces majority of research outputs (Gazni, Sugimoto and Didegah, 2012). Most collaboration studies use co-authorship relations and acknowledgements or sub-authorship (Cronin, Shaw and La Barre, 2003) from publication databases to investigate collaborative relations.

The results of successful research are usually published as scientific papers, thus if multiple researchers work together to produce new knowledge, it is likely that they will be co-authors in the published scientific papers. So even though collaboration is not always equivalent to co-authorship (Bar-Ilan, 2008; Katz and Martin, 1997), multiple-

authored papers are widely used to indicate the extent of collaboration between individuals or organisations, which is why the standard methodology for studies of collaboration in academia involves the analysis of co-authorship data obtained from publication databases like the Web of Science and Scopus or from a sample of a few core journals of a particular field (Katz and Martin, 1997; Newman, 2004; Glanzel and Schubert, 2005; Melin and Persson, 1996; Beaver, 2001). As collaboration is not always visible through co-authorship (Cronin, Shaw and La Barre, 2003), and a few core journals from a particular field may not reflect the collaboration patterns of the whole disciple (Beaver, 2001), Beaver (2001) advises that the results be qualified according to the data sources.

Gift authorship, where authors who may not have contributed to the work are included, and ghost authors, where authors who made significant contributions are not included in the published scientific paper (Cronin, Shaw and La Barre, 2003) are among several drawbacks to using co-authorship as an indicator for collaboration (Katz and Martin, 1997). However, the advantages of using this method include invariance, ability to be verified, relatively inexpensive, practical and the results being statistically more significant than those from case studies (Katz and Martin, 1997).

Since institutional collaboration decreases exponentially as the geographic distance separating collaborative partners increases (Melin

and Persson, 1996), collaboration is influenced by geography, as is university website inter-linking (Thelwall, 2002d).

Although data from publication databases are arguably the most reliable source that can be used to indicate the extent of collaboration between organisations, they are not publicly available, can be expensive to access and it is time consuming to process the author affiliation fields to extract institutional information, so studies have investigated alternate data sources like hyperlinks and organisation name mentions. This thesis investigates ways in which web based data like hyperlinks and organisation name mentions can be improved to be better used to study collaboration between academic institutions.

Hyperlinks in websites have some similarities with citations in academic published articles, which is why early webometrics studies emphasised the conceptual similarities between links and citations in articles (Rousseau, 1997). Relationships between links in websites are described in **Figure 2.1** (Björneborn, 2004):

Table 2.1 Description of linking relationship in **Figure 2.1** (Björneborn, 2004)
B has an in-link from A; B is in-linked; A is in-linking ; A is an in-neighbour of B
B has an out-link to C; B is out-linking; C is out-linked; C is an out-neighbour of B
B has a self-link; B is self-linking
A has no in-links; A is non-linked
C has no out-links; C is non-linking
I has neither in- nor out-links; I is isolated
E and F have reciprocal links; E and F are reciprocally linked
D, E and F have in- or out-links connecting each other; they are triadically interlinked
A has a transversal out-link to G: functioning as a shortcut
H is reachable from A by a directed link path
C and D are co-linked by B; C and D have co-inlinks
B and E are co-linking to D; B and E have co-outlinks
Co-inlinks and co-outlinks are both cases of co-links

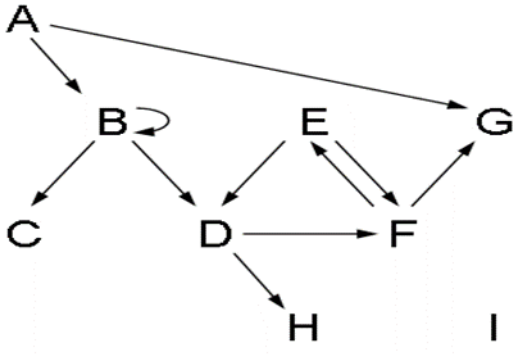

Figure 2.1 Basic link relationships between web pages A to I.

Early webometric studies involved applying bibliometric and informetric techniques to web based data, and particularly hyperlinks. Over the years, webometric studies have grown and do not simply involve applying bibliometric methods on web, and so Thelwall (2009) defined webometrics as *"the study of web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study".*

Webometrics cuts across multiple disciplines and it is related to web mining in computer science, in the sense that it also involves

analysis of web data. The main difference between webometrics and web mining is that webometrics is concerned with relating results to social science goals, but web mining research aims to identify methods that aid in efficient information retrieval or knowledge extraction from the web data.

2.2. Web mining

Web mining is the application of data mining techniques on the web (Etzioni, 1996). It uses methods from different research communities, such as information retrieval, information extraction, database technology and machine learning, to achieve its goals. In this respect, webometrics and web mining are similar because they both apply techniques from different research communities. The main difference between these fields is the final goal of the research. Webometrics is concerned with relating results found using different methods to offline occurrences; its ultimate goal is for social science reasons, while web mining focuses on solving information overload problems on the web; the ultimate goal is to improve users' experience as they interact with the web.

Web mining can be decomposed into four tasks (Kosala and Blockeel, 2000):

- Resource finding: retrieving documents from the web.
- Information selection/pre-processing: Removal of noise.
- Generalization: Pattern discovery among websites.
- Analysis: Interpretation of patterns.

Zhang and Segall (2008) added visualization to the list of web mining tasks. These sub-tasks are also carried out in webometrics research. Webometrics research needs to identify and retrieve web based resources or web data that can answer its research questions. This data is pre-processed to remove noise, for example Thelwall (2001a) restricted hyperlink data to those links that were from web pages related to research. Generalization attempts to identify patterns using different methods, which could be mathematical or statistical models (Payne and Thelwall, 2004; Vaughan and Thelwall, 2005). In webometrics, interpretation or analysis aims to find relationship between the web data identified and offline occurrences. For example, linking between universities has been shown to be largely influenced by the geographic distance separating these institutions (Thelwall, 2002d).

Even though it is hard to make a clear distinction between the types of web mining, there are three main categories: Web content mining, web structure mining and web usage mining (Kosala and Blockeel, 2000).

2.2.1. Web Content Mining

Web content mining involves the analysis of the content of web pages, often with text mining techniques like machine learning, natural language processing. Statistical methods are frequently applied to this area of web mining. These methods are sometimes used to achieve categorization of web pages through clustering or classification based on the web page content in order to improve information retrieval on the web.

2.2.2. Web Structure Mining

Web structure mining involves analysis of the web link structure in order to find patterns. Popular link analysis algorithms like PageRank (Brin and Page, 1998) and HITS (Kleinberg, 1999) use web link structure to rank web pages to help web users to find relevant results from their search engine queries more easily.

Graph and social network analysis (SNA) are sometimes used in web structure mining tasks because the web can be represented as a directed graph, where each web page is a vertex and two vertices are connected by an edge if there is a link between the two web pages. Barabasi and Albert (1999) showed that the distribution of degrees of the web graph structure follow a power law, and Broder and his

colleagues (2000) showed that the graphical structure of the web looked like a giant bow tie with a Strongly Connected Core (SCC) and two components on each side, one containing pages that can be reached from the SCC (In) and the other pages from which the SCC can be reached (Out). In addition there are two further collections of pages: Tendrils (indirectly connected to one or more of the above sets) and Disconnected (not connected to any of the above sets of pages).

2.2.3. Web Usage Mining

Web usage mining typically involves analysis of web users' interactions with the web. Data retrieved from server logs, click streams, and cookies are mined in order to identify users' behavioural or browsing patterns. The main goal of usage mining is to use the identified patterns to improve systems or the overall accessibility of websites. Usage mining is also used for business intelligence and mining web users' browsing patterns raises some ethical issues because users may have privacy concerns and may not consent to their usage data being used for marketing purposes.

2.3. Link analysis

Although the web is unstructured, hyperlinks not only aid navigation but can also indicate semantic relationships between web documents. This is exploited in web structure mining, such as through HITS (Kleinberg, 1999) and PageRank (Brin and Page, 1998).

Link analysis involves the study of link relationships between groups of websites or the link structure of a group of websites.

2.3.1. Academic web analysis

Link analysis has been extensively used in the study of academic web spaces. Web links can be studied with a micro, meso or macro approach (Stuart, 2008; Ingwersen and Björneborn, 2005). Micro analyses studies the inter page connectivity and what individual links may represent, while meso analyses also studies the inter site connectivity and macro analyses focuses on the connectivity between Top Level Domains (TLD) (Ingwersen and Björneborn, 2005).

Thelwall (2001b) showed that web links can be used to identify offline occurrences when his study showed evidence that the number of inlinks to an institution's web page in the UK correlates with research outputs and restricting the analysis to only those pages that were related to research increases the correlation (Thelwall and Harries,

2003). Inlink counts also correlate with the age of the website (Vaughan and Thelwall, 2003) which is not surprising as websites may get more inlinks as they become more well-known over time. The number of inlinks to computer science departmental websites in the UK correlates with research productivity, and their Web Impact Factors (WIF) correlate with their Research Assessment Exercise (RAE) ratings of computer science departments (Xuemei et al., 2003), however the correlation between RAE rating and library and information science departments was only limited (Thomas and Willett, 2000). The Web Impact Factor of a website (Ingwersen, 1998) is the total number of external links to a website divided by the number of web pages in the website. Xuemei and her colleagues (2003) used the number of staff members as the denominator as opposed to number of web pages in the calculation of WIFs in order to avoid penalising universities for publishing prolifically. Faba-Pérez et al. (2005) compared the value of different web indicators for a set of 1,180 websites through correlation, and found that there is no relationship between the WIF and Google's PageRank formula. Both WIF and Google's PageRank had no linear relationship with other web indicators.

Geographic distance, quality of university faculties and the language of a university were shown to affect link creation in Canadian Universities (Vaughan and Thelwall, 2005; Vaughan and You, 2009),

and the geographic distance separating universities also affect link counts in UK universities (Thelwall, 2002d).

Ortega et al. (2008) used hyperlink relationships between 535 universities from 14 European Union countries to analyse the topology of European academic network. They showed a visualisation of the social network relationship of academic institutions based on their links, in order to gain insights of the structure of the academic web space. The small world (highly clustered networks) properties of the hyperlink relationships between 109 UK universities have been investigated (Björneborn, 2006). Mathematical patterns have been identified in the link relationship between universities when Payne and Thelwall (2004) conducted a statistical analysis of the hyperlink structure of 111 UK universities, and Thelwall and Wilkinson (2003) analysed the mathematical graph structure of three national universities web spaces: Australia, New Zealand and the UK. Other researches have also carried out link based studies for institutions in countries like China (Meloche, 2010), India (Jalal, 2010), South Africa (Onyancha and Ocholla, 2007), Iran (Aminpour et al., 2009) and Nigeria (Nwagwu and Agarin, 2008).

National clusters are dominant in the hyperlink relationships among European universities, and this hyperlink data agrees with bibliographic data based on co-authorship production (Figuerola and Alonso Berrocal, 2013), although research has also shown that hyperlink structures do not always reflect collaboration (Shari, Haddow

and Genoni, 2012; Kretschmer, Kretschmer and Kretschmer, 2007). Kretschmer and her colleagues (2007) advice that methods that take into account the motivations for link creation should be used if hyperlinks are to be effectively utilised for collaboration studies.

Another popular webometric application area is in the ranking of world universities, because webometric indicators can reflect scientific activities, and rankings based on webometric indicators correlate with other non-web based ranking; Bibliographic Ranking, Times Ranking, Shanghai Ranking (Aguillo et al., 2006). The ranking scheme (Aguillo, Ortega and Fernández, 2008) uses four indicators: visibility (50%); size (25%); rich files (12.5%); and Google Scholar (12.5%). Visibility is the total number of external inlinks. Size is the total number of web pages, excluding rich files. A rich file is any web document in pdf, ps, doc and ppt format. The Google Scholar count is the number of documents from that university that appear in the Google Scholar database. The exact elements of this ranking scheme change over time, however.

The triple helix (university-industry-government) relations are important for investigating and modelling the relationships between knowledge and innovation (Leydesdorff and Etzkowitz, 1996; Etzkowitz and Leydesdorff, 1995). Analysis of triple helix relations is achieved with bibliometrics data; publications (Rafols and Meyer, 2010) and/or patents (Leydesdorff and Meyer, 2008; Meyer, Siniläinen and Utecht, 2003) even though publications and patents are inherently different

(Meyer and Bhattacharya, 2004). Although relations between university-industry-government can also be investigated with webometrics data (Minguillo and Thelwall, 2012; Priego, 2003).

Some other studies have attempted to identify the suitability of other web based data sources for webometric analyses. Evidence have shown URL or title mentions' network diagrams to be appropriate alternatives for hyperlink based networks (Thelwall, Sud and Wilkinson, 2012), impact studies (Thelwall and Sud, 2011) and web mentions can be used as substitutes for in-links in the Spanish academic web space (Ortega, Orduña-Malea and Aguillo, 2013).

This thesis investigates how machine learning methods can be used to improve the quality of webometrics data so it can be more suitable for academic web based analyses. In section 5.1.5, the extent to which title mentions are appropriate for collaboration studies is investigated.

2.3.2. Political web analysis

Hyperlinks have also been used to identify linking patterns among institutions or organisations. An analysis of the websites of 299 members of the 17th National Assemble of South Korea (Park and Thelwall, 2008) gave evidence of the existence of political agenda and affiliations with political parties in their linking patterns, but other offline

characteristics like demography or education were not identified. A study of a different topics showed that co-links between media and parties of the same political orientation in Spain are more common than between those with different political orientations (Romero-Frías and Vaughan, 2012) and web linking also show political biases in the European Union (Romero-Frías and Vaughan, 2012) while also following geographic patterns (Holmberg and Thelwall, 2009); from a study of interlinking of government websites in Finland.

2.3.3. Business web analysis

Link analysis can also be used as a source for business intelligence because the majority of links in commercial websites are created for business reasons (Vaughan and Gao, 2006) and the in-link counts to company websites, in some cases positively correlate with business performance in terms of revenue, profit and research (Vaughan and Wu, 2004; Vaughan, 2005). The correlation between in-links and some financial variables is also significant among organisations in the global banking industry (Vaughan and Romero-Frías, 2010).

Multi-Dimensional Scaling (MDS) for co-linked web pages has also been used to cluster similar companies, the MDS map showed a competitive landscape of the companies studied (Vaughan, 2005) because organisations with co-linked business websites are likely to be

competitors (Vaughan and Gao, 2006). Other clusters identified were as a result of linguistic and geographic factors (Vaughan and Romero-Frías, 2010), and it is possible to use this technique (MDS of co-linked websites) for analysis of non-business related organisations (Vaughan and Romero-Frías, 2012; Vaughan, Tang and Du, 2009).

MDS maps of organisations' websites that show the organisations' competitors are more accurate when text and co-links are used than when only co-links are used to construct the MDS map (Vaughan and You, 2009).

2.3.4. Link classification

Raw link counts can be unreliable for link analysis because links are prone to spamming (Smith, 1999), which is why there have been attempts to find alternatives to simple link counting (Thelwall, 2002c). Link counts also require both quantitative and qualitative analysis to determine if it can be used as a reliable indicator (Scharnhorst and Wouters, 2006).

Seeber et al. (2012) showed evidence that although the presence of links do not indicate the reasons why links may have been created or the relationship between the interlinked organisations, statistical analyses of the network ties between organisations show properties that are influenced by factors such as geographic distance, size of an

institution or research quality of an institution. The challenge of webometrics is then to identify, if possible, what subsets of links may be responsible for a particular factor. This may be achieved through hyperlink classification.

Attempts have been made to classify the links in a university's website but there is no consensus as to how links can be classified. No single link interpretation is perfect but there are two main approaches to hyperlink interpretation (Thelwall, 2006):

- Interviewing a random selection of link creators about why they created a link or
- Classification of a random selection of links in a way that is helpful to the research goals

Author interviewing may give more accurate results because link creators are responding to questions about why they created a link but classification of links is a more practical approach as it is unobtrusive, and also web masters may not remember why a link was created years ago. Classification of a random set of links has been used in a number of studies that analysed the academic web space (Bar-Ilan, 2004, 2005; Wilkinson et al., 2003; Thelwall, 2003; Chu, 2005).

Links in academic websites have been classified as research or non-research related (Thelwall, 2001b), substantive or non-substantive (Smith, 2003a) and shallow or deep (Vaseleiadou and van den

Besselaar, 2006). Thelwall (2001b) classified hyperlinks in web pages of UK academic institutions as research related or not research related based on the content of the target page, which he noted was a practical step. Although human classification is subjective, some general rules were created to for the classification process. For example, departments' homepages, staff research profiles and web pages of research groups were classified as research related while electronic journal pages were classified as non-research related because they were not necessarily created by authors within the hosting university. Research related links were found to correlate more highly than general links with average research ratings of UK institutions, justifying the classification-based filtering.

Wilkinson and colleagues (2003) studied 414 random links between UK academic institutions in order to identify the motivations for hyperlink creation. Even though individual links were investigated, the reason for link creation was determined using the source page and target page. They suggest that this approach is difficult because it is impossible to guess the motivation for link creation and in some cases there could be several motivations. Thelwall (2003) studied 100 random inter-site links from a UK university's website to the homepage of another UK university. He grouped web pages into four categories: *navigational*: a link created to direct uses to other non-subject specific information; *ownership*: links to partners, often in the form of a

clickable image of the university's crest; *social*: links to institutions of collaborating research groups; and *gratuitous*: links created without any specific motivation. Bar-Ilan (2004), in perhaps the most systematic study so far, classified the link, source page and target page from different aspects (link context, link tone and several other properties), in a case study of eight universities in Israel.

Manual classification of individual links is infeasible for large scale webometrics studies because of the size of university websites. An effective method for hyperlink classification or identifying the reasons for link creation is needed if links are to be harnessed to their full potential in webometric studies. Automatic classification of web pages with machine learning is a potential solution, this technique is used in Chapter 4 for automatic classification of university web page types. Machine learning has not been extensively used in webometrics research but has been applied to several computing based web studies (Chau and Chen, 2008; Luo et al., 2009; Qi and Davison, 2009).

2.4. Machine learning (ML)

Machine learning can be applied in webometric research, for example to identify characteristics that are unique to a particular type of link to automate the process of hyperlink classification. Machine learning is an area in computer science that deals with pattern discovery. *Supervised*

machine learning is concerned with teaching machines to classify or predict unseen cases of input data based on patterns identified from previously observed examples, usually called a training set. In contrast, *unsupervised* machine learning can be used for clustering, where nothing about the domain to be learnt is known. This section mainly describes the most widely used supervised and unsupervised learning algorithms without going in depth into the technical details of the algorithms.

2.4.1. Supervised learning

Supervised machine learning is also seen as classification. The goal of supervised machine learning is to identify a model that maps instances in a dataset to its respective class label. The assumption is that if the identified model accurately maps the known instances to their respective class labels, it will accurately predict the labels of new instances with unknown labels.

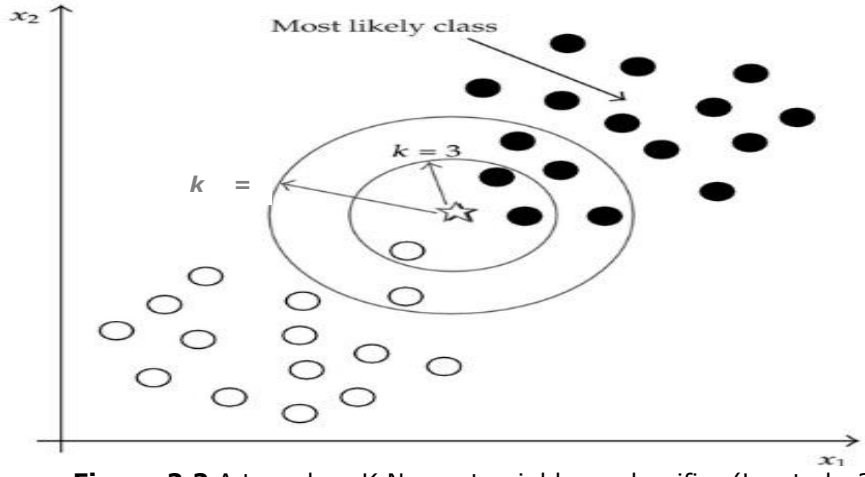

Figure 2.2 A two class K Nearest neighbour classifier (La et al., 2012)

There are several supervised learning algorithms. To apply these algorithms to webometrics efficiently, it is important to understand the algorithms so that the ideal technique that is suitable for the particular problem can be selected. This section of the thesis focuses on the supervised machine learning algorithms that are used in Chapter 4 to automatically classify academic web page types, and explains why they are suitable for these tasks.

2.4.1.1 K-Nearest neighbours

K-nearest neighbour classifiers (Cover and Hart, 1967) are instance-based learning algorithms called learning classifiers because they delay the generalization process until the classification is performed (Kotsiantis, 2007), and thus requires less computational time in the training process. Instance-based classifiers have an advantage over other learning algorithms because of their simplicity. The K-nearest neighbour algorithm has just one adjustable parameter, k, which controls the number of nearest neighbours that are used to define class membership.

Figure 2.2 shows the schematic of a two class K-nearest neighbour classifier. When the instance (designated by star) is classified with either K = 3 or K = 7, it is assigned to the black class because more instances in the most similar 3 and 7 neighbours belong to the black class. Similarity between neighbours can be computed with a variety of metrics, the most popular of which are listed in

Table 2.2. These metrics can also be used to determine the quality of clustering solutions.

A drawback of instance-based classifiers is the computational time that it takes for classification. The time it takes to classify an instance is proportional to the number of training instances and the number of features that describe each instance (Guo et al., 2003). Instance-based algorithms like the k nearest neighbour classifier are stable, in some cases, the classification accuracy does not drop significantly when up to 80% of training instances are removed (Kotsiantis, 2007). Because it only has one parameter "k", it may be easy to find the optimal classification models, even though there is not a principled way to choose "k" (Kotsiantis, 2007), only through methods like (Guo et al., 2003) that increases the already poor computational time.

Table 2.2 K Nearest neighbour distance metrics (Kotsiantis, 2007)

Minkowsky	$D(x,y) = \left(\sum_{i=1}^{m} \lvert x_i - y_i \rvert^r\right)^{1/r}$
Manhattan	$D(x,y) = \sum_{i=1}^{m} \lvert x_i - y_i \rvert$
Chebychev	$D(x,y) = \max_{i=1}^{m} \lvert x_i - y_i \rvert$
Euclidean	$D(x,y) = \left(\sum_{i=1}^{m} \lvert x_i - y_i \rvert^2\right)^{\frac{1}{2}}$
Camberra	$D(x,y) = \sum_{i=1}^{m} \frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert}$
Kendall's Rank Correlation	$D(x,y) = 1 - \frac{2}{m(m-1)} \sum_{i=j}^{m} \sum_{j=1}^{i-1} sign(x_i - x_j) sign(y_i - y_j)$

2.4.1.2 Decision tree induction

Utgoff (1989) formally defined decision trees as "*a leaf node (answer node) that contains a class name, or a non-leaf node (decision node) that contains an attribute text with a branch to another decision tree for a possible value of the attribute".* Decision tree induction recursively splits data into disjoint sets according to a criterion. Each node is a feature that an instance can have, and leaf nodes contain output classes for instances that reach that node. Classification of instances starts at the root node, and then the instances traverse down the tree in the direction that meets several criteria until a leaf node is reached. The value of the leaf node is then assigned to that instance.

Constructing optimal binary decision trees is a NP-complete problem (Kotsiantis, Zaharakis and Pintelas, 2006), however several techniques, like the C4.5 algorithm (Quinlan, 1993) and CART, an acronym for Classification And Regression Trees (Breiman et al., 1984) can be used to build decision trees. The C4.5 algorithm is implemented in the popular machine learning toolkit, WEKA (Hall et al., 2009) that is used in this thesis to automate the classification of academic web pages in Chapter 4. The J48 algorithm in the WEKA machine learning toolkit is an implementation of the C4.5 decision tree induction algorithm.

Two major phases of some types of decision tree induction algorithms are the growth phase and the pruning phase (Kotsiantis,

2013). The growth phase involves splitting the training data into disjoint sets and the pruning phase reduces the size of the decision tree to avoid overfitting. Pruning is necessary to reduce the size of an over grown decision tree and also to avoid overfitting (Bradford et al., 1998).

Decision tree induction Pseudo Code (Kotsiantis, 2007)

```
1 For each attribute a
2    Find the feature that best divides the training data
3 Let a_best be the attribute that best splits data
4 Create a decision node that splits on a_best
5 Recurse on the sub-lists obtained by splitting on a_best
  and add nodes as children until leaf nodes contain only
  one instance
```

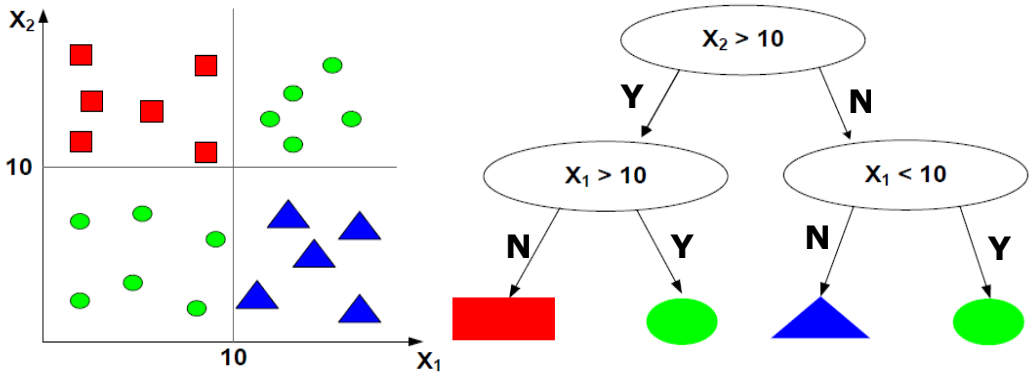

Figure 2.3 An example of a decision tree classifier

A major difference between the C4.5 and CART algorithms is the way that the best feature that separates the training data is selected. The attribute with maximum information gain is used to split the training set. C4.5 uses entropy to compute the information gain, while CART uses the Gini index. The entropy is calculated by:

$$Entropy(S) = -\sum_{i=1}^{n} Freq(C_i, S) * log\big(Freq(C_i, S)\big)$$

$Freq(C_i, S)$ is the relative frequency of instances in class C_i.

The Gini index is computed by:

$$GiniIndex(S) = 1 - \sum_{i=1}^{n} Freq(C_i, S)^2$$

And information gain is computed by:

$$InformationGain(S, A) = I(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * I(S_i)$$

The formula above computes the information gain of attribute A in data set S where $i = 1 \dots n$ are possible values of attribute A, $S_i \dots S_n$ are partitioned subsets of S where attribute A is i , $I(S)$ is the entropy in the C4.5 algorithm and the Gini index in the CART algorithm.

The advantage decision trees have over other supervised learning algorithms is that the resulting tree can be seen as a set of rules which makes it easy to understand the classification model.

2.4.1.3 The Perceptron Algorithm

The perceptron algorithm (Rosenblatt, 1962) is one of the simplest classification algorithm. The simplest classification problems are dichotomous. The input x can belong to one of two classes $(y = 1 \; or \; y = -1)$. For a given input represented as a vector, $x = (x_1 \dots x_n)$, the goal of

machine learning is to classify x to either of two classes $(-1, 1)$. For a given input x, linear models describe a function $g(x)$ that returns the output y based on a linear function determined from a training dataset $D = (x_1 y_1), \ldots, (x_n y_n)$. Linear discriminant functions for pattern classification is described in (Highleyman, 1962).

The perceptron algorithm (Rosenblatt, 1962) is a linear classifier written as:

$$g(x) = sign\left(\sum_{i=0}^{d} w_i x_i\right); \quad x_0 = 1$$

Here, x is the input vector and the goal of the perceptron algorithm is the find the appropriate values of the weight vector w that accurately classifies the training dataset.

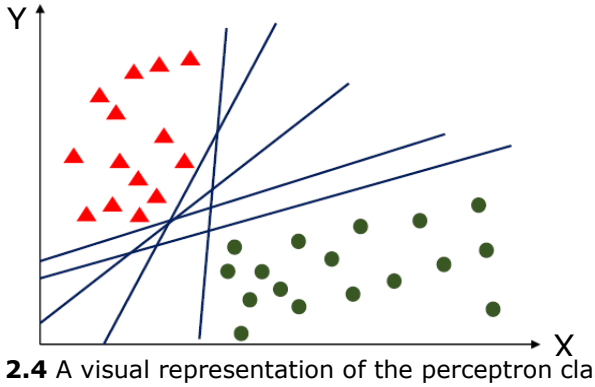

Figure 2.4 A visual representation of the perceptron classifier

The perceptron algorithm aims to find any line that can separate the two classes in a dataset. It finds this line by iteratively updating the

weight vector until all objects in the dataset are correctly classified, if possible. The perceptron learning algorithm is mathematically proven to get a solution if the input space is linearly separable (Rosenblatt, 1962).

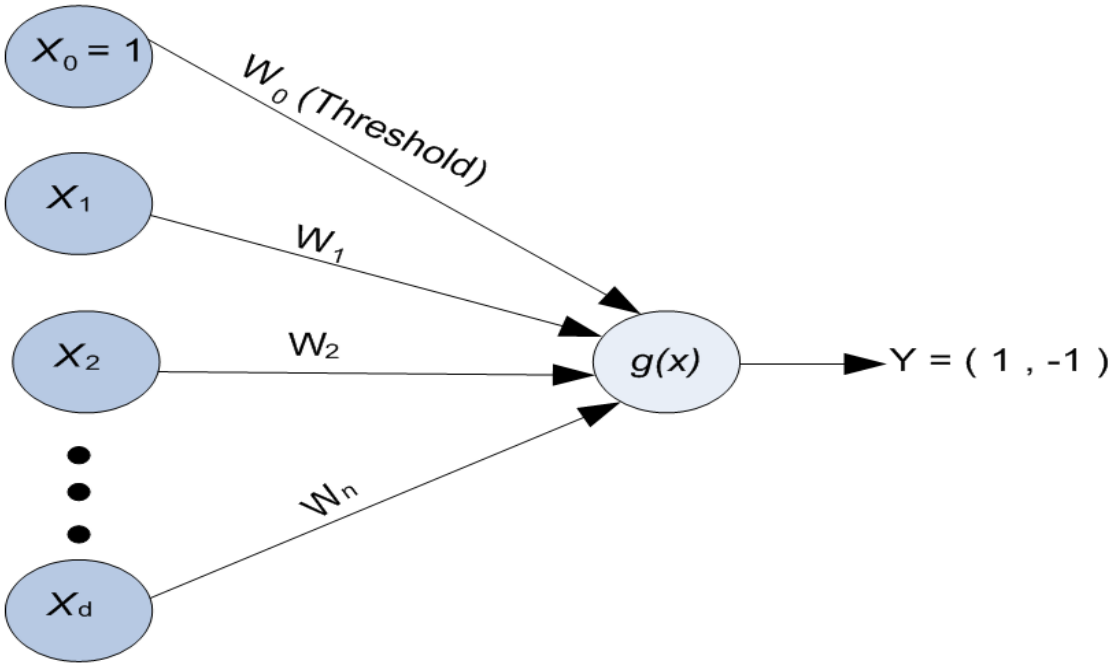

Figure 2.5 The perceptron linear classifier

Perceptron like algorithms may have advantages when data has a high number of features with few relevant ones (Kotsiantis, 2007). However, a simple perceptron algorithm will not find an adequate solution if the data is not linearly separable and most real world machine learning problems are not linearly separable.

2.4.1.4 Multilayer Perceptron

Multilayer perceptron (Ruck et al., 1990) is a kind of neural network. Neural networks are models based on the human nervous system. Essentially, they are black boxes that are able to predict the output

class when they recognize a given input pattern. Single perceptrons will not reach a solution when the training set is not linearly separable. This can be overcome by using artificial neural networks (Rumelhart, Hinton and Williams, 1986), that combine multiple linear classifiers, hence the name multilayer perceptron.

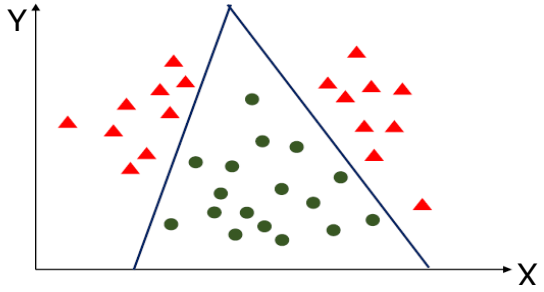

Figure 2.6 Multiple layered perceptron

Multilayer perceptrons are made up three components, the input, hidden layers and the output. The feed forward nature of the multilayer

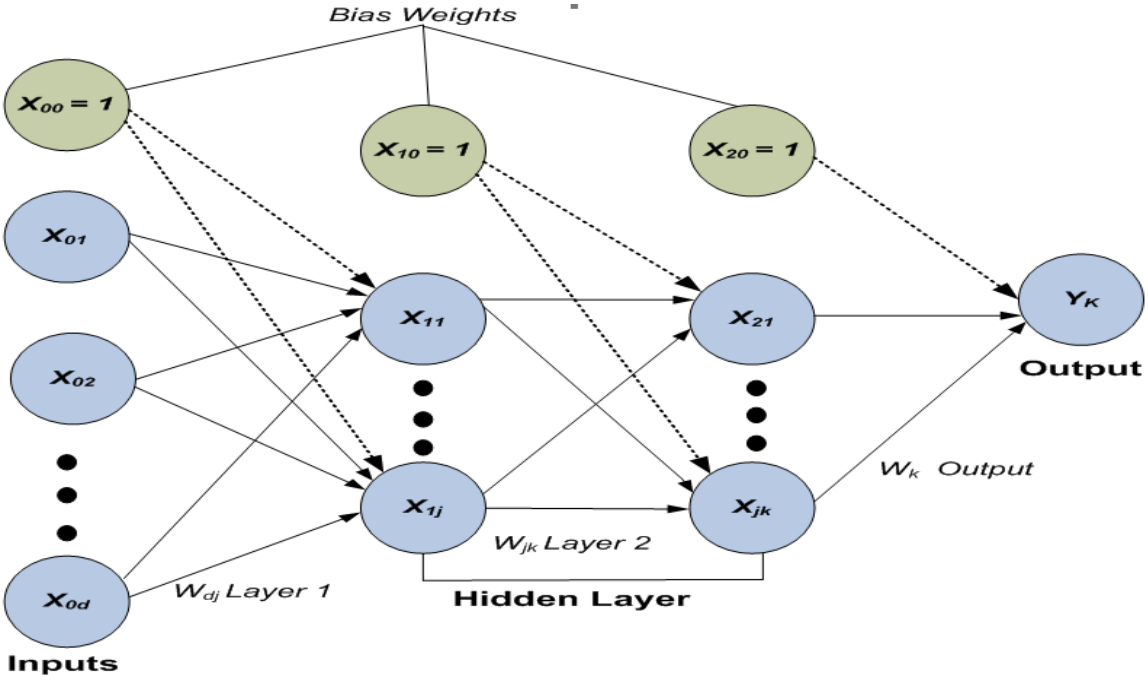

Figure 2.7 Feed forward artificial neural networks

perceptron means that the output of the previous layer is the input of the next layer. The back propagation algorithm (Rumelhart, Hinton and Williams, 1986) is widely used to determine the weights of perceptrons in each layer of the neural network. Training of neural networks can be improved by pruning (Castellano, Fanelli and Pelillo, 1997; Parekh, Yang and Honavar, 2000); which is removing redundant neurons or weight vectors.

One of the main challenges of the multilayer perceptron is finding the optimal number of layers. Underestimation can lead to poor generalization while overestimation can lead to overfitting, also training with the back propagation algorithm may find solutions slower that other machine learning algorithms.

2.4.1.5 Naïve Bayes

The Naïve Bayes algorithm can be defined as *"a classification algorithm based on Bayes rule that assumes the attributes X_1 ... X_n (input sample) are all conditionally independent of one another, given Y (class label)"* (Mitchell, 2010).

Naïve Bayes classifiers are statistics-based learning algorithms. Bayes classifiers use conditional probabilities of random variables in their algorithms. For example, the probability that it is raining and cloudy P(x = Raining | y = cloudy) is higher than the probability that it is raining and sunny P(x = Raining | y = sunny).

Classification with the naïve Bayes algorithm is determined by the formula:

$$P(class|input) = \frac{P(input|class)P(class)}{P(input)}$$

This essentially is the same as maximizing the numerator because $P(input)$ is the same for all classes. The probability of class $P(class)$ is computed by the number of training examples in class c divided by the total number of training examples.

$$P(class) = \frac{N_c}{N}$$

To estimate the conditional densities $P(input|class)$, the naïve Bayes classifier assumes that all parameters of the input vector are conditionally independent. If the input is represented as a feature vector of $input = P(x_1, x_2, \ldots x_n)$, $x_1, \ldots x_n$ are conditionally independent. A way to estimate a class conditional density is using a multinomial class conditionally density. As all input features are assumed to be

conditionally independent given the class, $P(input|class)$ is simple the product of the probability of each input feature given the class.

$$P(x_1, x_2, \ldots x_n \,|class\,) = \prod_{i=1}^{n} P(x_i|class)$$

The naïve bayes algorithm's computational time for training is significantly shorter than other machine learning algorithms (Kotsiantis, 2007).

2.4.1.6 Support vector machines (SVM)

Like all supervised machine learning techniques, the goal of support vector machines is to identify patterns in a training set and then use

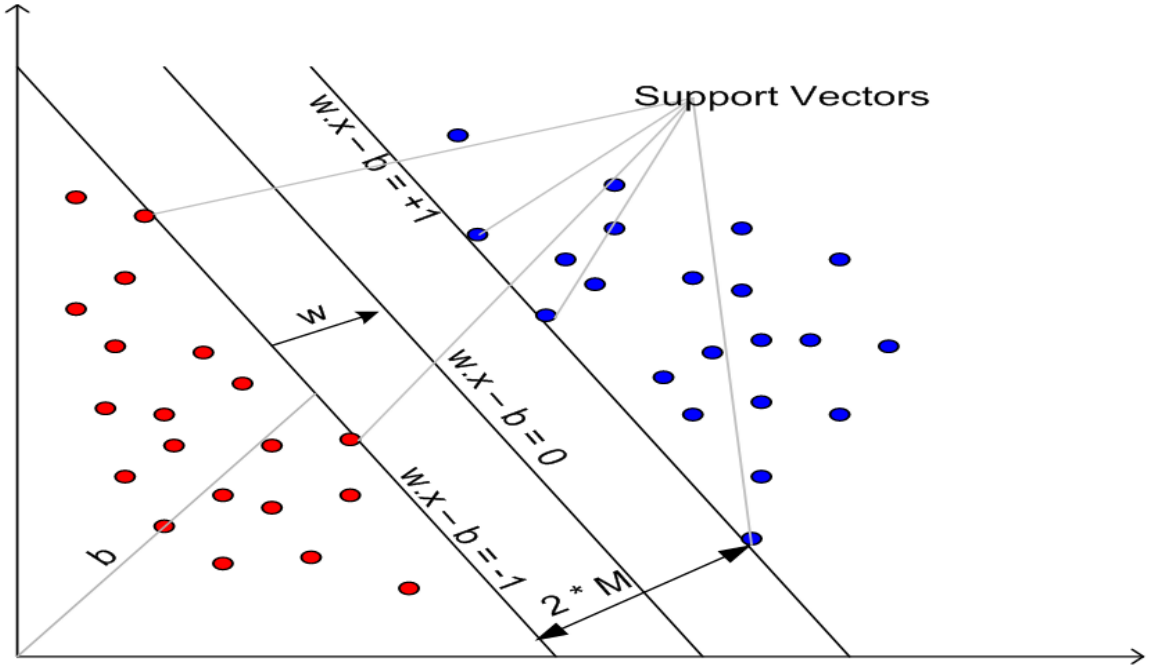

Figure 2.8 A graphical representation of support vector machines.

these identified patterns to predict future unseen cases. Support vector machines (Cortes and Vapnik, 1995) do this by describing a hyper plane that separates the training data whist maximizing the distance separating the two classes.

Figure 2.8 is a graphical representation of support vector machines. The SVM aim is to identify the hyper plane "$w.x - b = 0$" that separates the classes in the training set by the distance 2M, where M is the maximum distance from any class to the separating hyper plane. The input data on the margins act as support, and hence are called support vectors, and only these points are required to predict future classifications which make the data required for classification much smaller than the size of the training set.

The task of finding the optimal separating hyper plane can be mathematically represented as a dual optimization problem and solved using quadratic programming.

2.4.1.6.1 Mathematical representation of SVMs

From geometry, the distance between two hyper planes; $wx - b = -1$ and $wx - b = +1$ is $2/\|w\|$. Support vector machines aim to maximize this margin whilst meeting the constraints that:

$$w.x - b \leq -1 \; when \; class \; label \; y_i = +1 \text{ and} \qquad 2.1$$

$$w.x - b \geq +1 \; when \; class \; label \; y_i = -1 \qquad 2.2$$

If equations 2.1 and 2.2 are combined to one equation, it results in:

$$y_i(w.x_i - b) - 1 \geq 0 \qquad 2.3$$

Hence the mathematical problem of SVMs is to:

$$\begin{aligned} &Minimize \; \frac{\|w\|^2}{2} \\ &Subject \; to \; y_i(w.x_i - b) - 1 \geq 0 \end{aligned} \qquad 2.4$$

This optimization problem can be solved using Lagrangian multiplier method. The SVM problem in Lagrangian is:

$$\begin{aligned} \zeta(w, b, \alpha) = \frac{\|w\|^2}{2} - \sum_{i=1}^{n} \alpha_i \left[y_i \left(w.x_i \right) - 1 \right] \\ \alpha_i \geq 0 \quad \forall \, i \end{aligned} \qquad 2.5$$

When the derivative of $\zeta(w, b, \alpha)$ with respect to w and b is computed, it implies that:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \; ; \sum_{i=1}^{n} \alpha_i y_i \qquad 2.6$$

Substituting the weight vector "w" back to the Lagrangian equation (2.5) results in the SVM dual optimization problem:

$$\begin{gathered} Maximize \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (x_i, x_j) \\ Subject\ to \quad \alpha_i \geq 0 \quad \forall\, i \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{gathered} \qquad 2.7$$

The conversion of the original SVM problem (2.5) using Lagrangian multiplier to the SVM dual optimization problem (2.7), makes the training data (x_i, x_j) occur only as dot products. This product is usually substituted though kernel functions. For example, in WEKA (Hall et al., 2009) there are several implementations of kernel functions that can be used to determine an SVM learning model. Using different kernel functions may result in better or poorer learning accuracy, depending on the machine learning problem.

Research shows that support vector machines can perform either as well or significantly better than competing methods in most machine learning contexts (Burges, 1998) and so are a logical first choice for webometrics. However, the no free lunch theorem (Wolpert and

Macready, 1997) suggests that no algorithm can outperform all others in all machine learning problems. So even though SVMs most times outperform other machine learning algorithms, it is worth investigating if any algorithm will be best suitable for a particular learning problem.

2.4.2. Unsupervised learning

Supervised learning, as discussed earlier, often involves classification, where all categories are where all categories are known beforehand. The task of classification is to identify a function to identify a function that will differentiate future examples based on patterns learned from a patterns learned from a training set. In unsupervised learning, however, the output categories the output categories are not known beforehand and therefore this type of learning is often of learning is often exploratory in nature. Unsupervised learning is often used for clustering. used for clustering. Clustering is defined as: "*given a set of n objects, find k groups based on a find k groups based on a measure of similarity such that similarities between objects in the between objects in the same group are high while objects in different groups are low"* (Jain, *groups are low"* (Jain, 2010). The similarity between objects can be computed using the K computed using the K nearest neighbours distance metrics in

Table 2.2. Clustering has been applied to a variety fields, from medicine to economics. It has also been applied to webometrics (Thelwall, 2002b), to study the link relationships between UK university websites using different statistical techniques. The results suggested that the web links in UK academic websites can be mined, but only with suitable methods if complex patterns are to be extracted. This thesis

uses machine learning methods to extract complex patterns (collaboration) from web data in UK academic websites.

2.4.2.1 Traditional clustering techniques

Although there are thousands of clustering algorithms (Jain, 2010), early clustering algorithms are broadly divided into two types: partitional and hierarchical.

2.4.2.1.1 Partitional clustering

The objective of partitional clustering is to split a set of n objects into a set of k clusters in order to minimize an objective function. Usually, this objective function is the minimum square distance. The most common partitional clustering algorithm is the **k-means** algorithm (Macqueen, 1967). The goal of the k-means algorithm is to minimize the sum of squares error over all clusters. The k-means algorithm finds a partition such that the square error between the cluster centres and the objects in that cluster is minimized.

$$SquareError = \sum_{k=1}^{k} \sum_{x_i \in c_k} \| x_i - \mu_k \|^2$$

Here, k is the number of clusters, x_i is an object, c_k a cluster and μ_k the centre of clusterc_k. K-means is considered to be *greedy* in that, it iteratively assigns an object to its closest cluster, which may not be the

best choice in the long run. As a result of this, it tends to converge to a local minimum. The best result is achieved when the algorithm is run several times with random initial cluster centres. Minimizing the objective function of k means has been proven to be an NP-hard problem even for k = 2 (Drineas et al., 2004).

The basic k-means algorithm is (Tan, Steinbach and Kumar, 2005):

```
1 Select K points as initial centroids
2 Repeat
3   Form K clusters by assigning each point to its closest
  centroid
4   Recompute the centroid of each cluster
5 Until centroids do not change or repetitive cycles
  emerges
```

There have been several variants of k means. ISODATA (Ball and Hall, 1965) initialises with a high number of clusters and then automatically identifies the optimal number of clusters by merging clusters whose similarity between their centres are higher than a specified threshold. Fuzzy c-means (Dunn, 1973) and improved by Bezdek (1981) allows objects to belong to more than one cluster. The degree of object membership to a cluster is between 0 - 1.

2.4.2.1.2 Hierarchical clustering

Hierarchical clustering can either be agglomerative or divisive. In agglomerative mode, the algorithm goes from the bottom upwards. Each object to be grouped starts as a single cluster, and the algorithm goes upwards, grouping smaller clusters to form a larger one. Divisive mode goes from the top down, starting with one cluster and recursively dividing large clusters into smaller clusters. When there is a set of n objects to be clustered, the input of a hierarchical clustering algorithm is usually an *n x n* similarity matrix. The output of a hierarchical clustering algorithm is a sequence of nested clusters, with each cluster being a partition of the set of objects. Most hierarchical clustering algorithms are agglomerative (Tan, Steinbach and Kumar, 2005) and follow a basic algorithm:

```
1  Assign objects to a cluster
2  While number of clusters is greater than one
3      Find the closest pair of clusters and merge them
4  Return a sequence of nested clusters
```

Computing the similarity/closeness between pair of clusters can use several methods. Common methods are:

Single-linkage: merges two clusters based on the most similar objects among all clusters; *Complete-linkage:* merges two clusters based on the least similar objects among all clusters; *Average-linkage*: merges clusters based on the average similarity between each object in a cluster and all objects in each of the other clusters; *Median*: the pair

with the closest median or cluster centre is merged; *Ward:* The smallest mean square distance when pair of clusters are merged.

Unlike partitional clustering, hierarchical clustering does not need to specify the number of classes beforehand, or to set initial conditions (e.g., cluster centres or thresholds). However, when clusters are merged or divided, they cannot be undone. In partitional clustering, such as k means, data may move from a cluster to another if it is closer to the new cluster's centroid. The advantages of partitional clustering are the disadvantages of hierarchical clustering and vice versa.

2.4.2.2 Graph-based clustering

A graph is a structure formed when a set of vertices V are connected by a set of edges E. A graph could be directed or undirected, weighted or un-weighted. If the objects to be grouped are represented as a graph, the goal of graph clustering is to group vertices in such a way that there are many edges within each group and few between groups (Schaeffer, 2007). This makes graph clustering particularly suitable for networked data, because networked data can be represented as a graph, and a visualisation of this graph or clusters formed from the analysis of the graph can give insights into the connectivity of entities (e.g. webpages) on the graph in order to identify central/important actors, which may be difficult to identify with other clustering techniques. For this reason, graph-based clustering techniques have been widely used in webometric

studies. Software like Pajek (de Nooy, Mrvar and Batagelj, 2005); that implements a number of graph based techniques, is a tool used in a number of webometric studies (Ortega and Aguillo, 2008; Thelwall and Zuccala, 2008; Holloway, Bozicevic and Börner, 2007; Ortega and Aguillo, 2009; Holmberg and Thelwall, 2009; Leydesdorff and Vaughan, 2006) for data analysis.

Numerous graph clustering algorithms exist, but early algorithms were based on the minimum spanning tree (MST), graph partitioning or highly connected sub graphs, these algorithms are implemented in social network analysis software packages like Pajek (de Nooy, Mrvar and Batagelj, 2005).

2.4.2.3 Self-organising maps (SOM)

A widely used clustering technique based on neural networks is the Kohonen Self-Organising Maps (SOM) (Kohonen, 1990). SOM is an abstract mathematical model of the visual sensors in the human brain (Yin, 2008). It uses a principle called competitive learning that results in the spatial organisation of the data to be analysed. Although SOM has a number of application areas such as; vector quantization (Heskes, 2001), data compression (Amerijckx et al., 1998), data visualization (Vesanto, 1999), classification (Lau, Yin and Hubbard, 2006) and

clustering (Vesanto and Alhoniemi, 2000). In Chapter 6, SOM is used for clustering and visualization of computer science research groups in the UK. SOMs result in a visualization of the input data where objects similar to each other are placed closer together on the map while objects less similar are placed farther apart.

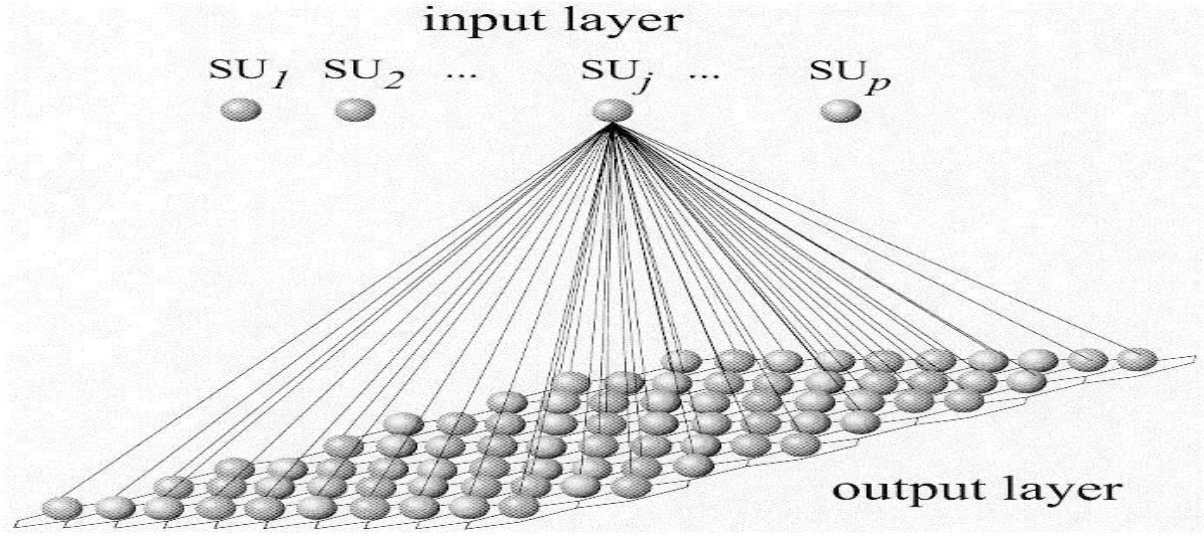

Figure 2.9 A visualisation of a two dimensional SOM (Giraudel and Lek, 2001)

Figure 2.9 shows the working principle of SOM. The high dimensional input ($SU_1 \dots SU_p$) in mapped into a two dimensional grid. The input is a vector that contains characteristics of each sample in the dataset. They act as a stimulus to neurons in the output layer (two dimensional grid). The SOM is formed in four stages:

- **Initialization:** The simplest initialization method is initializing all neurons on the output grid to random values, however other

techniques based on the Eigen values of the input vector can be used (Ballabio, Vasighi and Filzmoser, 2013).

- **Competition:** Each neuron in the output layer uses a discriminant function like the Euclidean distance to determine the neuron closest to a sample introduced to the grid. Samples are usually introduced sequentially or in a batch. The closest neuron on the output grid to the introduced sample emerges as the winner.
- **Cooperation:** The winning neuron then determines a topological neighbourhood of neurons close to it, based on a mathematical function.
- **Adaptation:** The value of topological neighbours of the winning neurons are then updated, based on how close they are to the winning neuron.

The spatial organisation of the output layer after all objects in the dataset are fed creates a visual representation of clusters in the dataset.

A matlab implementation of SOM (Vesanto et al., 1999) is used in Chapter 6. In section 6.1.1.1.1, input parameters of SOMs are described, as well as quality measures for self-organising maps.

2.5. Summary

This chapter gave a brief review of webometrics application areas and the machine learning techniques that will be used to analyse web data in subsequent chapters. Webometrics research initially involved applying bibliometric and informetric research methods to web because of the conceptual similarity between citations in peer reviewed journal articles and hyperlinks in web pages. Over the years, the methods used in webometrics research have cut across multiple disciplines which is why a more recent definition emphasises the social science goal of webometrics research and understates the restriction of using informetric methods for analysis.

Techniques used in webometric research are also used in a computer science discipline, web mining. Web mining is the application of data mining techniques on the web. These two research areas, webometrics and web mining share similar sub tasks, like finding web resources, pro-processing the data to remove noise, discovering patterns and analysing patterns. The difference between these research areas is that web mining is aimed at improving users' experience as they interact with the web, while webometrics is more concerned with relating web behaviour to offline occurrences.

Link analysis has been successfully used in a number of webometric studies. Correlations between inlinks to a university and the university's research productivity have been often studied. Link analysis can also be used to identify political trends and for business intelligence. Inlinks to a company's website correlate with business performance and multi-dimensional scaling of co-linked business web pages show a landscape of a company's competitors.

In spite of the potential of hyperlink analysis, links are unreliable and prone to spamming. Because of this, a number of studies have tried to understand the meaning of hyperlinks or aimed to identify the reason why hyperlinks are created in web pages. Current methods used to analyse hyperlinks involve the manual classification of links in a way that is helpful to the research goals but the web is huge and growing exponentially so manual link analysis is becoming more and more impractical. It has been recommended that if links are to be effectively used for webometrics research, a way in which hyperlinks can be automatically classified in necessary.

Automatic classification of hyperlinks can be achieved using machine learning. Supervised learning techniques can be used to learn patterns and automatically classify hyperlinks, which will be a step in the right direction for harnessing the full potential of hyperlinks in websites. Several studies have already attempted to classify hyperlinks in academic web pages but no consensus has been reached as to how

the classification process can be used effectively in large scale. So automating a link classification scheme increases the possibility for more effective webometrics research.

The result of data clustering can be used to identify behaviour for better insights into a particular domain. Clustering techniques have already been applied to webometrics research but the methods are largely graph theoretic based. Other unsupervised learning techniques may be able to identify behaviour not already known through graph clustering methods.

3. Webometrics research methods

This chapter describes the main methods used in the empirical chapters of this thesis. It starts with describing quantitative and qualitative research. This thesis investigates how web data can be used to study collaboration between UK universities. From the definition of webometrics in (Thelwall, 2009), it can be suggested that this thesis is largely webometric research because web data is analysed using machine learning methods to investigate collaboration, which essentially is a social science goal.

Webometrics research is primarily quantitative but has some qualitative aspects. For example, the motivations for link creation can be identified by qualitative research methods, that is, interviewing link creators about the reasons why they may have created a link in their web sites (Thelwall, 2006), so section 3.1 discusses quantitative and qualitative research methods.

Section 3.2 and 3.3 reviews the data collection techniques used in webometrics studies; by web crawling and by search engines, and then the next sections in this chapter describe how statistical and machine learning methods can be used in the analysis of the kind of web data used in the empirical chapters.

3.1. Quantitative and qualitative research

Webometrics studies predominantly use quantitative research methods to find any association or relationship between different variables. Quantitative research is objective and seen as hard science while qualitative research methods are seen as soft science, and qualitative research often answers its research questions by analysing what people say in interviews or focus groups in order to find peoples' opinions about a particular subject. **Table 3.1** from (Anderson, 2006) lists some of the differences between quantitative and qualitative research methods.

Table 3.1 Difference between quantitative and qualitative research methods (Anderson, 2006)

Quantitative	Qualitative
Objective	Subjective
Hard science	Soft science
Test theory	Develops theory
One reality: focus is concise and narrow	Multiple realities: focus is complex and broad
Measurable	Interpretive
Reports statistical analysis.	Reports rich narrative, individual interpretation
Basic element of analysis is numbers	Basic elements are words and ideas
Researcher is separate	Researcher is part of the process
Establishes relationship, causation	Describes meaning, discovery
Strives for generalization, generalization leading to prediction, explanation and understanding	Strives for uniqueness. Patterns and theories developed for understanding
Highly controlled setting: experimental setting. (Outcome oriented)	Flexible setting. Process oriented

"Quantitative research is concerned with the collection and analysis of data in numeric form. It tends to emphasize relatively large scale and representative sets of data. Qualitative research is concerned with collecting and analysing information in as many forms, chiefly non numeric. It tends to focus on exploring, in as much detail as possible, smaller number of instances which are seen as interesting or illuminating, and aims to achieve depth rather than breadth" (Blaxter, Hughes and Tight, 1996).

3.2. Data collection by search engines

Search engines are primarily designed to aid users retrieve information from the web. Popular search engines like Yahoo!, Google and Bing crawl the web and store web pages in a database. Relevant web pages are then presented to a user based on the user's query and the search engine's retrieval algorithm. Although the main purpose of a search engines is web search, search engines can be used to gather data for quantitative webometrics research. The hit count estimate, which is the number of results that match a query, can be used as the raw data in webometrics studies (Kousha and Thelwall, 2007).

Search engine queries could be keyword searches or link/domain name searches, which make it possible for the hit count to be used as an estimate of the number of links or co-links between websites. Hit counts have been used to identify countries or organisations with highest web impact (Ingwersen, 1998), to retrieve co-link counts of business websites (Vaughan and You, 2008) and several other studies (Hoerlesberger and van den Besselaar, 2003; Thelwall, Vann and Fairclough, 2006).

Applications Programming Interfaces (APIs) for search engines allow software like Webometric Analyst 2.0 (http://lexiurl.wlv.ac.uk/) to send automated requests, and download the results of search queries.

There are drawbacks to using hit count estimates as the raw data for webometrics research, because not all web pages are indexed by search engines (Lawrence and Giles, 1998, 1999; Smith, 2003b; Bar-Ilan, 2002) and even pages in a search engine's database that match a query may not be in the search results (Bar-Ilan and Peritz, 2004). Thelwall (2008b) lists possible causes of this as errors in the search engine's parser program, the maximum number of results reached (1000 as at June 2007) or databases not fully searched as a result of insufficient time. The effect of incomplete results can be reduced by splitting search queries into several disjoint sub-queries (Thelwall, 2008a). Also, Thelwall (2008a) advises that search engines should not be expected to deliver correct results but should be seen as engineering products that provide relevant results for their users. Commercial search engines determine which web pages are relevant to their users based on complex heuristics or algorithms that are not always available to the public, and the features of search engines change frequently so it is difficult to explain any inconsistency that may exist in results.

The limited correlation between the link counts and the RAE or library and information science departments in (Arakaki and Willett, 2008) was attributed to the nature in which link data can be collected

from search engines, which suggests that it may be difficult to achieve adequate results for departmental level webometrics studies with search engines as the primary source for data collection.

Using search engines as a primary source for data collection for some webometric research is the most practical or sometimes the only feasible data collection method, especially if the goal of the research is to identify trends on the whole web. If the study involves a small set of websites, a personal web crawler may be a more reliable solution for data collection.

3.3. Data collection by web crawling

Extracting information from websites can be achieved through web crawling. Web crawling is important for webometrics research because several search engine features that can be used to for data collection in webometrics research data are being deprecated or have reduced functionality.

The web can be seen as a large graph where web pages are represented as nodes and hyperlinks between these pages are edges. A web crawler traverses this graph structure systematically whilst extracting and saving the required information from the web pages.

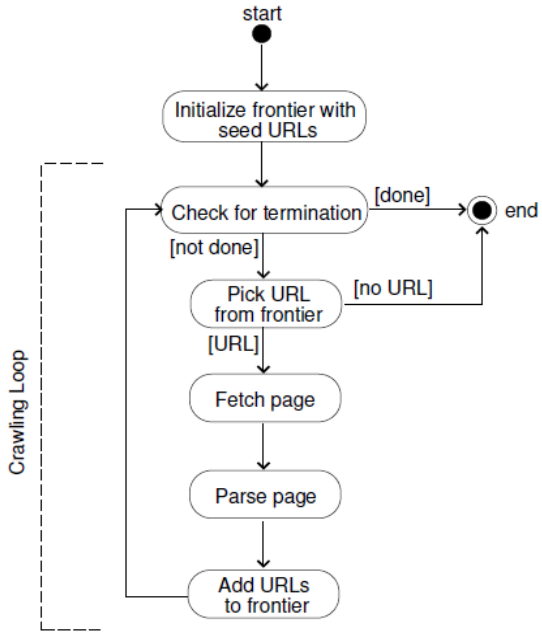

Figure 3.1 Flow chart of a typical web crawler (Pant, Srinivasan and Menczer, 2004)

Figure 3.1 is a flowchart showing the procedure of a typical web crawler. The crawler starts by adding URLs (Seed URLs) to initialize the frontier: an initial list of web page URLs that should be visited by the crawler. Then the crawler checks if the stopping condition is met. This is an important step because the web is huge, so crawling can sometimes take up to several months or longer. An appropriate terminating condition is needed to stop the crawling process as soon as sufficient data has been collected. Common stopping conditions are the maximum number of pages or the maximum depth of a website that the crawler can visit. Next, a URL is selected from the frontier. The simplest selection method is a FIFO (First in First Out), which selects the oldest link that was added to the frontier. FIFO essentially gives a breadth first search of the web graph from the seeds. The crawler then retrieves the web page that the selected URL points to, then extracts and saves the

necessary information and the links that will be added to the frontier. This process is repeated recursively until either the terminating condition is met or the frontier is empty.

The basic crawling procedure is simple but Page and his colleagues (1999) showed that designing a web crawler for a search engine is a complex task and the most fragile component of a search engine. A common problem is identifying the most appropriate URL to add to the frontier.

A FIFO approach may be appropriate for webometrics link data collection, because in most cases the link data needed is only from a particular a group of websites. For more effective web crawling however, different metrics should be used to add or remove web page URLs from the frontier. For example, when unlimited time or resources are not available for crawling, it is necessary to systematically order URLs so that the crawler will visit important web pages significantly faster than a simple FIFO approach can (Cho, Garcia-Molina and Page, 1998). Cho, Garcia-Molina and Page (1998) listed some crawling metrics:

- Back-link Count: URL ordering based on the number of in-links to the web page that the URL points to. Web pages with more in-links are visited first.

- PageRank Score: Unlike back-link count, which treats all links as the same, PageRank ranks web pages based on their popularity. Links from more important web pages have higher weights than links from less important web pages.
- Forward-link Count: This metric uses the number of links in a web page to determine the web page's importance. Cho, Garcia-Molina and Page (1998) argue that web pages with many links are important because they may be directories.
- Location: Web pages closer to the homepage are assumed to be more important that web pages farther from the homepage. Distance from the homepage can be approximated by the number of slashes in the URL.

In cases where the goal is to crawl web pages of a particular theme, the crawling metrics listed in (Cho, Garcia-Molina and Page, 1998) are not ideal. Focused crawling (Chakrabarti, Berg and Dom, 1999) adds only those URLs relevant to the theme or topics to the crawler frontier. Metrics like the best first search (Menczer et al., 2001) use the cosine similarity between web pages and a specified query/string that is represented as a word vector to determine which web pages should be visited next. Shark Search (Bra and Post, 1994) use the link anchor text and the text surrounding links to compute the cosine similarity of a URL with a query/string. Classification techniques can also be used for focussed crawling (Chakrabarti, 2003), categories

and examples of web page and relevant topics are previously provided by the user, and then a Bayesian classifier that can computes the probability that a URL is of a particular theme is built, based on the training examples.

Boilerplate are those sections of web pages that are identical across the majority of web pages in a website (Baroni and Kilgarriff, 2006), these among others include web page headers, web page footers, navigation bars, menu items, privacy notice and advertisements (Marek, Pecina and Spousta, 2007). So if the web pages crawled are to be used for any kind of textual analysis, it is important to exclude this noise (boilerplate). There are techniques that can be used to remove boilerplate from web pages (Marek, Pecina and Spousta, 2007; Kohlschütter, Fankhauser and Nejdl, 2010). However, boilerplates does not greatly affect the crawled data used in the empirical chapters. In Chapter 4, only the text wrapped in the HTML <title> tag is used for classification and in Chapter 6 text data is filtered to contain only a predetermined set of phrases, this therefore reduces the effect of boilerplates.

3.3.1. Crawling ethics

Crawlers that are designed for search engines traverse the whole web, unlike crawlers that are designed for webometrics research that will

retrieve data from a single website or group of websites. As a result of this, webometric crawlers tend to send multiple requests to a single server which could result in denial of service if not controlled. Denial of service causes a web server to crash or slow down significantly because it is flooded with more requests than it can handle. Thelwall and Stuart (2006) suggest the following ethical guidelines for crawling websites.

- Websites should not be crawled unless it is absolutely necessary. Data available from other sources like commercial search engines or previous crawls may be sufficient for the research needs. For example, (Cybermetrics, 2011) has a comprehensive link structure of several world universities.
- Be aware of the financial implications crawling may cause to website owners and should be prepared to compensate website owners if necessary.
- Consider privacy implications. Websites usually have web pages that should not be visited by web crawlers. A list of these web pages is contained in the robot.txt file.
- Follow robots.txt guidelines
- Be polite and not flood websites with too many requests in a short space of time.

3.4. Statistical analyses

If web based data is a reliable indicator for investigating collaboration between organisations, this data should have statistical relationship with other data sources traditionally used to study collaboration. Quantitative statistical data analyses are used to investigate the association between web data and traditional data that is used to study collaborative relationships. Quantitative statistical data analyses techniques are applied in Chapter 5.

Webometric studies are concerned with relating web based data with offline occurrences. For example, in order to investigate if there is collaboration between web actors in web based data, the association between web based data and data traditionally used to study collaboration should be analysed. Co-authorship relations are seen as the de facto standard to measure collaboration, so if there is an association between web based data and co-authorship data, then this suggests that web based data may be an alternative data source for collaboration studies. The extent to which two variables are related can be estimated through statistical correlations, but results must be interpreted with caution because correlation does not imply causation. Even though the correlation between two variables is high, it may because of some unknown causes, correlation is a resultant of the

influence of all related factors (Wright, 1921), both known factors and unknown factors.

The Pearson correlation coefficient indicates the relatedness of two continous variables. Pearson's r can be between -1 and +1. The closer r is to -1 or +1 indicates the level of association between the two variables. r = -1 suggests a negative relationship, and that the data lies on a straight line with a negative slope, R = 0 means that there is no linear relationship between the two variables and r = 1 suggests a positve linear relationship between two variables. When the distribution of data is not normal, for example bibliographic and hyperlink data do not follow a normal distribution, Spearman's rank correlation should be used instead of Pearson correlation because Pearson's correlation is sensitive to outliners and assumes that the data is normally distributed. Spearman's rank correlation is less suseptible to outliners because it uses the ranking of the variables instead of the actual value of the variables to compute correlations.

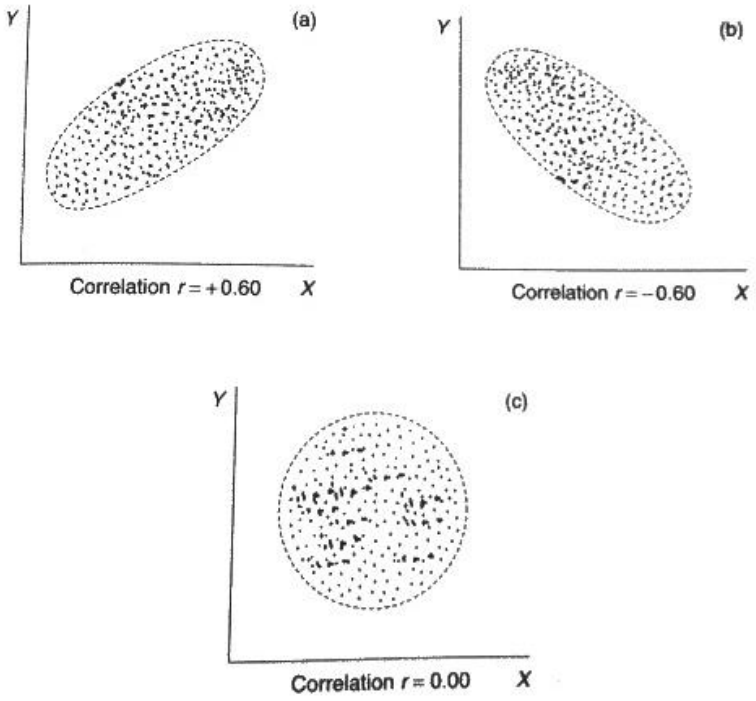

Figure 3.2 Scatter plot of two variables and their corresponding correlation coefficient (Jaeger, 1990)

Figure 3.2 shows a scatter plot of between two variables. When the correlation between the variables are 0.6 and -0.6 the scatter plot shows a linear relationship between the two variables but when the correlation is 0, there is no visible relationship between the variables on the scatter plots, the points on the map are in random locations.

3.5. Exploratory data analysis

Webometrics research is sometimes exploratory, thus data clustering techniques that can be used to explore data for knowledge extraction is used to analyse UK computer science research groups in Chapter 6.

Cluster analysis can be used to identify previously unknown characteristics that may be present in web data. Factor analysis is a statistical technique that is used for dimension reduction. Factor analysis reduces the number of variables by combining correlating variables into a common factor (a principal component). The new reduced variables that are derived from the originals summarize the information contained in the original variables. Factor analysis is implemented through principal components analysis (PCA) in the SPSS statistical package. Even though the main objective of PCA is dimension reduction, Ding and He (2004) have shown evidence that the principal components or factors can be interpreted as solutions of the k means objective function, so it can be easily adapted for clustering purposes. Other clustering algorithms that were discussed in section 2.4.2 can also be used for exploratory analyses of web data. Self-organising maps are used for exploratory web data analysis in Chapter 6 because unlike other clustering techniques, the clustering solution is a topological ordering of samples to be clusters, where similar clusters are placed closer together and less similar clusters are placed farther apart. In chapter 6, the PCA components are also used in a simple clustering algorithm to determine if PCA components can produce suitable clustering solutions, compared to self-organising maps**.**

3.6. Automatic classification

The supervised learning (classification) algorithms that were described in section 2.4.1 are used in Chapter 4 and 5, in an attempt to improve the extent to which hyperlinks can be used to investigate the extent of collaboration between universities.

Supervised learning follows the procedure in Figure 3.3 for automatic classification. First, an appropriate classification algorithm that can successfully classify an input data set should be selected. When the appropriate learning algorithm has been selected, characteristics or features of each instance in the dataset must be chosen. Certain features may be dependent on another, thus removing these redundant features or reducing the dimension of instances can improve the speed as well as the accuracy of the learning algorithm. There are many different feature selection techniques (Yu and Liu, 2004; Markovitch and Rosenstein, 2002) that ranks features based on their importance. In Chapter 4, the information gain is used to rank features because the decision tree algorithm used, uses the information gain to determine the features in each node on the tree. Feature selection often occurs after algorithm selection because some algorithms use particular type of features. For example, support vector machines work with real valued features. If the SVM is the algorithm of choice, then it is essential that all features are real values and so

discrete or categorical variables should be converted to continuous variables. In chapter 4, the supervised learning techniques that were described in the literature review (see section 2.4.1 page 37) were compared to determine the most suitable for the particular machine learning problem in this thesis (classification of academic web page types).

The model selection or training phase is the step where parameters of the learning algorithm are tuned to generalize the input data set. The model is then evaluated in the testing phase (Training and Validation). If the accuracy is determined to be sufficient, depending on the learning problem, then the model is used for classification, otherwise, the learning process is repeated from one of the previous steps.

Poor results may occur because the input data is imbalanced, a less optimal learning algorithm or feature set was used, or because the learning model over-fits the training set. Overfitting is when the learning model performs well in training but much worse in the test phase.

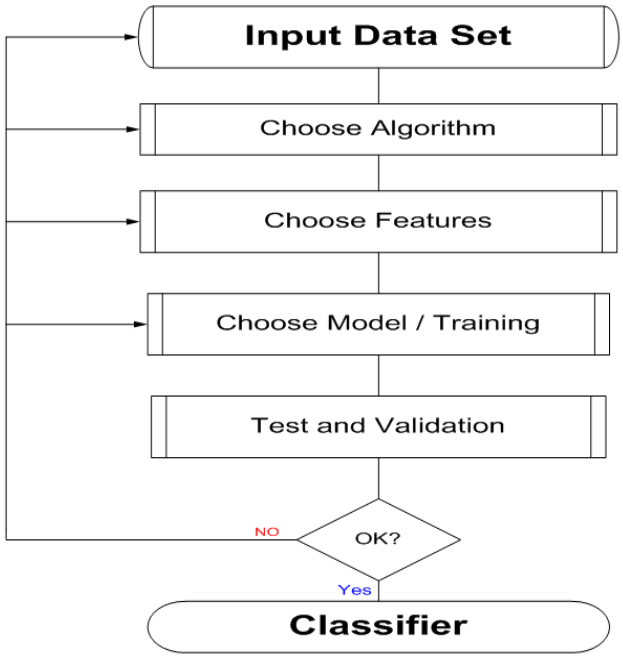

Figure 3.3 Flow chart of a typical supervised learning procedure

3.6.1. Testing and evaluation

A supervised machine learning algorithm trains itself on a training data set and evaluates its accuracy on a test data set. The fundamental assumption of machine learning is that the distribution of training data is representative of the distribution of test data and future examples (Liu, 2006). If the learning algorithm accurately classifies the test set, then the machine learning assumption suggests that it will perform well for future unseen cases. Accuracy is measured with the equation:

$$Accuracy = \frac{Number\ of\ correctly\ classifed\ examples}{Total\ number\ of\ examples}$$

Ultimately, the minimum acceptable level of accuracy depends on the application the learning model is designed for. The researcher has to decide the level of error that is acceptable for their study.

Precision, recall and F-measure can also be used to estimate the performance of a learning algorithm. They are calculated based on four parameters: True Positives (TP), False Negatives (FN), False Positives (FP) and True Negatives (TN). Given a test set D, assume that each instance of the set can be of class $y = (1, -1)$ and that $f(x)$ is the function trained to predict future unseen instances. The parameters TP, FN, FP and TN are:

- True Positives (TP) = $Count(f(x) = 1 \; and \; y = 1)$
- False Negatives (FN) = $Count(f(x) = -1 \; and \; y = 1)$
- False Positives (FP) = $Count(f(x) = 1 \; and \; y = -1)$
- True Negatives (TN) = $Count(f(x) = -1 \; and \; y = -1)$

Precision, recall and F-measure are computed by:

$$Precision = \frac{TP}{TP + FP}; Recall = \frac{TP}{TP + FN}; \mathrm{F} = \frac{2 * Precision * Recall}{Precision + Recall}$$

In n fold cross validation, the input data is divided into n equal disjoint subsets. Each subset is used as the test set, and the union of the others the training set. The accuracy of the classifier is then the average accuracy of the n different subsets.

3.7. Summary

This chapter described the main techniques that are used in the studies carried out in subsequent chapters. Data for webometrics studies can be extracted from search engines, individual websites through web crawling or information extraction. Using search engines for data collection is the most practical solution for data collection in some types of webometrics studies but it has its drawbacks, for example not all web pages are indexed by search engines and even when pages are in a search engines database, it may not be listed in the search results. The effect of incomplete results can overcome by query splitting. However, the procedures that commercial search engines use to determine relevant results are not always available to the public. Using personal web crawlers to collect data is an alternative if the webometrics study involves a small set of websites.

The simplest method of web crawling is a breadth first search of the web graph, where new web pages to be visited are added or removed from the frontier in a FIFO approach, but more complex approaches, like focused crawling, are needed for effective and more ethical web crawling.

Crawling websites poses some ethical concerns like the cost to web page owners, privacy and denial of service. These ethical issues

must be addressed before a website is crawled. Websites should not be crawled unless it is absolutely necessary and all other alternatives for data collection have been exhausted. Web crawlers should also be designed in such a way that unnecessary requests are not sent to web servers.

More efficient link analysis can be achieved when links are previously automatically classified with supervised learning techniques. Supervised learning techniques can also be used to automatically process web page URLs before they are added to the frontier for more efficient focused crawling, and to determine if a web page will likely contain data needed for the research. This way, some unnecessary requests to web servers can be avoided.

Finding relationships between web based data and offline phenomena can be achieved through statistical data analysis. For example, Pearson and Spearman correlations can be used to estimate the relatedness of two variables. Other behaviour can be investigated through exploratory cluster analysis. Although there are numerous clustering techniques, SOM is used in Chapter 6 because of the topological ordering of input samples the clustering solution results in, and a simple factor analysis may be easily transformed to a clustering solution as it has been shown that factors in a principal component analysis can be seen as solutions to the k means objective function.

4. Automatic classification of university web page types for large scale webometrics studies

This chapter addresses the first aim of the thesis, which is identifying an effective way in which the reasons for hyperlink creation in academic websites can be identified. It describes an automatic classification-based method that can be used to suggest the reasons why a hyperlink in academic websites has been created, thus increasing the possibility of fully harnessing the potential of hyperlinks in webometrics research.

The reason why a link has been created in a website can be inferred by studying the two pages that the hyperlink connects. If university web pages are grouped into categories, then the relationship between the aggregate of links connecting two categories may also reflect that of the links from individual web pages. Answering the following research questions will help to reach the goal of automatic classification:

1) How reliably can machine learning techniques classify university web page types?

2) How reliably can machine learning techniques predict the classification of link target pages from characteristics of link source pages?

3) What are the common characteristics of out-links from each university web page type?

Section 4.1 describes how the data used in this study was collected through web crawling. A classification scheme is described in Section 4.1.1, procedure for using machine learning techniques to classify academic web page types or predict the target web page type of a link is described in Section 4.1.2 and 4.2.1, and then the result of manually investigating the reasons for outlinks in a sample of links from the different web page types is presented in Section 4.2.2. Section 4.3 summarises the findings from this study.

4.1. Methods

A custom web crawler was designed to retrieve the link structure of 111 UK universities. The crawler extracted links originating from a UK university to another UK university, although not visiting all pages in a university's website. It only covered those web pages that could be reached by iteratively following links from a university's homepage, similar to SocSciBot (Thelwall, 2003). SocSciBot was not used in this thesis because it limits crawling to a maximum of three depths, which was not suitable because, as this research is concerned with only links between UK academic websites, the web crawler should retrieve as many links as possible, regardless of the depth, whilst stopping the

crawl when there seems to be no additional link to another UK university website. Also, SocSciBot does not allow the download of additional information from websites. For example, the web page title may be useful for automatic classification, but it is not possible to retrieve this information using SocSciBot.

As this study is only concerned with hyperlinks between UK academic institutions, an additional constraint that new web pages should not be added to the frontier (list of websites to visit) when the crawler visited 2000 consecutive pages without finding a link to another UK university was added to the crawling algorithm. The number 2000 was set heuristically to ensure that the majority of web pages in a university website were visited whilst crawling was stopped in a reasonable amount of time.

Pseudo code to crawl a university website

```
1 Add homepage (seed URL) to the frontier
2 Repeat
3     Get web page at the head of the frontier
4     Extract all URLs from web page
5     Add all URLs to the link structure file
6     If not reached 2000 consecutive web pages with
  .ac.uk TLD
7         Add all URLs in the same domain as the
  university domain to the frontier
```

```
8  While frontier is not empty
```

4.1.1. Page Types

Higher Education Institutions have three main missions: teaching, research and the "third mission". García-Aracil and Palomares-Montero (2009) stated that this third mission comprises of *"entrepreneurialism, innovation and social commitment"*.

If university websites are designed to channel the activities and functions of a university, which in turn are in line with the three main missions of HEIs, then it can be argued that most universities' websites could be identical in terms of the types of pages contained. In some cases, the text and structure of the homepage replicates the physical structure of that organisation (Weninger, 2012). Similar organisations arguably have identical web site structures. For example, computer science departments have identical website structures, with homepages containing links to people, research and courses (Weninger, Johnston and Han, 2013).

In this study, web pages are grouped into categories that are in line with the three missions of HEIs because most university websites are designed to channel their activities which are in line with these missions, and then the classification scheme is automated with supervised machine learning techniques.

There are a number of classification schemes for university web page types or hyperlinks, some of them were described in the literature review (see Section 2.3.4 page 33). However, most of the classification schemes discussed earlier were designed in a way that helped answer the research questions their research asked. Bar-Ilan's (2005) classification scheme is the most detailed and perhaps the most adequate, but because the classification scheme is multi-faceted, it will be difficult to automate with machine learning. Bar-Ilan (2005) classified links in four facets: the relationship between the source and target web pages, the reason for link creation, the tone of the link (positive, negative or neutral) and placement of link (part of a list or in the menu/sidebar). This study for simplicity infers the reason for link creation by investigating the most common reasons from subsets of links in the same web page category.

This research uses a classification scheme that is less detailed but simpler than a previous one (Bar-Ilan, 2005). A simple single faceted classification scheme will be easier to automate with machine learning and may achieve higher accuracy than a more complex multi-faceted one. However, the final web page types in **Table 4.1** to some extent overlaps with Bar-Ilan's (2005) identified reasons why web masters (creators) create web pages in academic websites. The web is constantly evolving and may have changed since (Bar-Ilan, 2005), also

the study was based on a case study of Israeli academic websites which may be different from UK academic websites. So a different page type classification scheme was developed as opposed to using Bar-Ilan's (2005).

The initial categories were *"teaching"*, *"research"* and *"business and innovation"*, from the three main missions of HEIs. Then, a random sample (100) of the university web pages was given to an independent researcher to manually classify. The description of the web pages identified by another researcher who is not the author of this thesis is in **Figure 4.1**. A total of 100 web pages are enough to give a broad idea about the web page categories. For example, a total of 100 random links that appeared in the homepages of UK university websites have previously been used to investigate the reasons for academic hyperlink creation (Thelwall, 2003). The descriptions of the 100 web pages were then grouped into the three main categories. If a web page that did not fit into any of the three main categories was found, a new category was created. **Table 4.1** reports the final web page categories found and their descriptions.

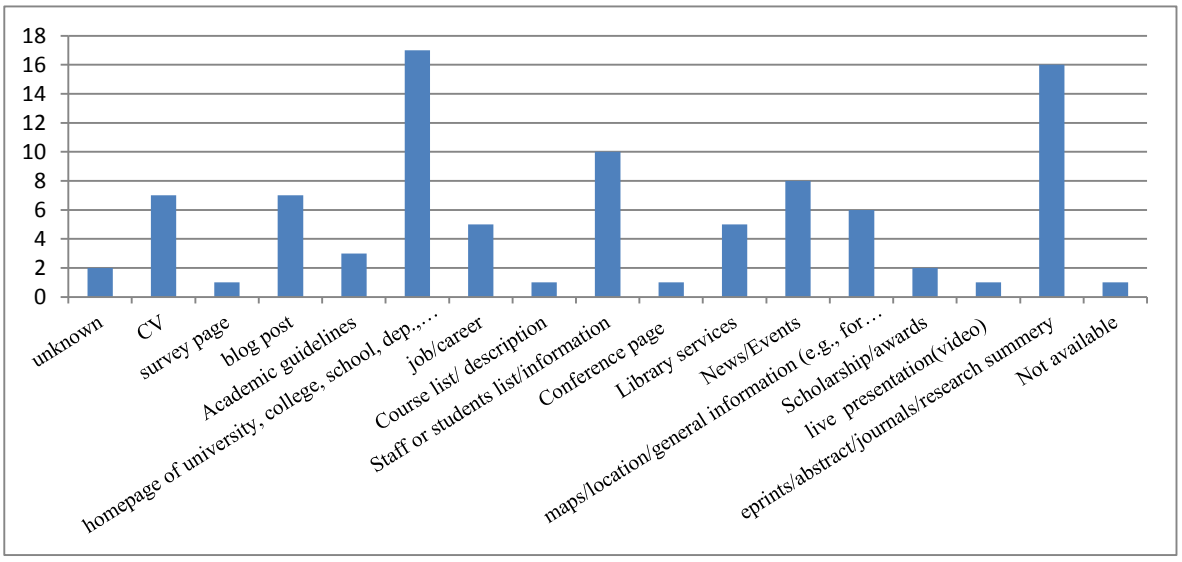

Figure 4.1 Description of the web pages pointed to by 100 randomly selected hyperlinks.

When the final categories were determined, a total of 2,549 links, which is approximately 0. 12% of the total number of unique links (2,148,979) in the 111 UK universities' websites were randomly selected. The web pages that these links pointed to were downloaded and manually classified by the author of this thesis into one of the final categories in the classification scheme.

The final class a web page belonged to is largely based on the opinion of the manual classifier (author of thesis). When a web page did not seem to fit in any of the categories it was assigned to the best fit, also in some cases some web pages could have been assigned to multiple categories. Effort was made to avoid creating new categories and assigning web pages to multiple categories. Assigning labels to multiple categories results in a more complex machine learning

problem. The traditional supervised learning algorithms described in the literature review (section 2.4.1) cannot effectively be used for multi label learning (Zhou and Zhang, 2006).

Table 4.1 Description of web page categories found in UK university websites and distribution of 2,549 manually classified web pages into categories.

Page Type	Number of web pages	Description
About	247	Promotes the school and gives information to staff/students. Examples of such pages are news, information, university profile, prospectus, events
Business and innovation	12	Connects the school to non-academic environment. Examples include expert services offered, community projects, partnership with science parks
Discussion	239	Forums, blogs or web page containing opinions of a user. Comments or posts in these pages are for a variety of reasons: research, teaching or recreational.
Support	890	Contains a repository of learning resources for students/staff support, skills for learning, services, counselling. Examples include archives, books, and database.
Research	729	Involved with the production of new knowledge. Examples include Research centres, research groups, research projects, academic schools/departments, conferences, abstracts and academic articles.
Staff	358	Related to a staff member in the university. Examples include staff homepages, staff profiles, lists of publication and CVs.
Student Life	45	Enhances the student experience. Examples include student union website, student benefits, campus facilities, tourism and recreation
Study	29	Involved with transfer of knowledge. Examples include module learning materials, module timetables, module page and lectures.

4.1.2. Automatic classification

The manually classified web pages formed the training data set that was used to automate the classification scheme. Classification of web pages by a single researcher makes the result subjective and 0. 1% of the total web pages is low, however manually classifying 2,549 web pages is time consuming, particularly if this task is done by a single person. A larger dataset may produce more accurate results but the current dataset may be sufficient to train a supervised learning classifier for the purposes of this study.

To ensure the validity of a classification scheme, inter coder agreement between multiple coders should be high. However, using multiply coders is not always possible (Thelwall, 2003). Simple and detailed classification schemes may not need multiple coders (Bar-Ilan, 2005) because categories links could belong to are less debatable. These types of classification schemes will have very high inter coder agreement. The webometrics studies reviewed in (Holmberg, 2009) achieved between 70% and 98% inter coder agreement when manually classifying hyperlinks.

These 2,549 pages (training dataset) were used to construct classification models and the accuracy of the classification models was determined using a 10 fold cross validation.

The features of each web page used for machine learning were derived from the web page title and/or web page URL, pre-processed and then represented as an inverse document frequency multiplied by the term frequency (TFIDF) vector or a binary vector.

Word Tokenization: Splits attributes (web page title/URL) into word tokens which could be unigrams and bigrams. For example "the quick brown fox jumps over the lazy dog" has 16 tokens: [the, quick, brown, fox, jumps, over, lazy, dog, the quick, quick brown, brown fox, fox jumps, jumps over, over the, the lazy, lazy dog]. A simple way to achieve tokenization is by assuming [space] separates word tokens. In this study, only words that did not contain any non-alphabetic characters were used and these characters "[space]\r\t\n.,;:\'\"()?!-><#$\\%&*+/@^_=[]{}|`" were used as word token separators.

Capitalisation: All characters were converted to lower case.

Stop word removal: Removal of the most frequent words that occur in the English language. Words like the, and, a, is ... are all removed. WEKA (Hall et al., 2009) contains a list of stop words. The 111 university names, their domain names as well as www, http and https were added to the list of stop words.

Stemming: Stemming reduces inflected words to their root form or stem. For example jumps and jumping have the same stem, jump. Accuracy may be improved if all words are represented in

their root form. The Porter Stemmer algorithm is one of the most commonly used stemming algorithm, it is used in this study. Stemming algorithms occasionally make errors. For example, the Porter Stemmer stems both university and universe to univers. This is called over stemming, both university and universe and stemmed to the same root when they should not be. Under stemming occurs when words that should have the same root when stemmed are stemmed to different roots. The majority of stemmers are prone to either over stemming or under stemming, which is why improving the precision and recall of stemming algorithms is still an open research area (Jivani, 2011).

The number of features will increase as the number of web pages in the training set increases. Too many irrelevant features can affect the speed and accuracy of a learning algorithm, so input features have to be carefully selected. The J48 algorithm (decision tree induction) uses the information gain metric to select the best features that optimally split the training data, so it is logical to use information gain to filter out some redundant features. The number of features influences the overall accuracy of a classification algorithm. This is evident in **Figure 4.2** where very high number of features negatively affected the accuracy of the classification algorithm. On average, the top 500 features performed well for all the classification algorithms in this study's

learning problem, so top 500 features were used to construct the classification models for the different supervised learning algorithms.

4.2. Results

How the features are pre-processed can influence the accuracy of the classification algorithm. **Table 4.2** shows the effect various pre-processing options have on the accuracy of the classifiers when automatically assigning web pages to the page categories in **Table 4.1**.

Table 4.2 A comparison of the accuracy of 10 pre-processing options for decision tree induction, support vector machines, k nearest neighbours, Naïve Bayes and a 3-layered neural network supervised learning classifiers for classifying the page types of 2,549 manually classified university web pages with baseline accuracy of 34.9%.

Bigrams /Unigrams	**TFIDF used rather than binary**	**Stem ming**	**Stop words removed**	**Page title included**	**URL included**	**DT Accuracy**	**SVM Accuracy**	**KNN Accuracy**	**Naïve Bayes Accuracy**	**MLP Accuracy**
*Unigrams	Yes	Yes		Yes	Yes	72.70%	78.30%	64.30%	72.70%	72.00%
Bigrams + Unigrams	Yes	Yes	Yes	Yes	Yes	72.60%	76.90%	67.70%	72.80%	68.60%
Unigrams		Yes		Yes	Yes	73.30%	76.80%	66.10%	69.80%	67.20%
Unigrams	Yes			Yes	Yes	72.30%	76.10%	64.10%	71.90%	69.40%
Bigrams + Unigrams		Yes	Yes	Yes	Yes	72.30%	75.90%	65.60%	70.80%	60.10%
Bigrams + Unigrams			Yes	Yes	Yes	71.70%	75.70%	65.30%	71.80%	65.00%
Unigrams	Yes	Yes	Yes	Yes	Yes	71.80%	75.50%	66.60%	71.70%	66.30%
Unigrams				Yes	Yes	71.50%	75.40%	65.00%	69.10%	63.50%
Unigrams		Yes	Yes	Yes	Yes	72.60%	74.10%	64.80%	68.60%	66.00%
Unigrams			Yes	Yes	Yes	72.60%	73.10%	65.10%	69.50%	60.90%

*Best pre-processing option and machine learning algorithm

Baseline accuracy is the accuracy of a classification model if all web pages are classified into the category with the most web pages. As expected, SVMs outperformed other supervised learning algorithms by up to 5% in terms of overall accuracy. **Figure 4.2** shows how the number of features affects the accuracy the classifier. The best pre-processing settings from **Table 4.2** were tested for different feature sizes. Initially, as the feature size increased, the accuracy of the classifier also increased but at some point the increase in size did not improve the accuracy much but reduces the speed of the classification algorithm for little accuracy gain. K-nearest neighbours and the multi-layered perceptron are the most susceptible to changes in the number of features. The accuracy of the classification algorithm was poorer when many features were used.

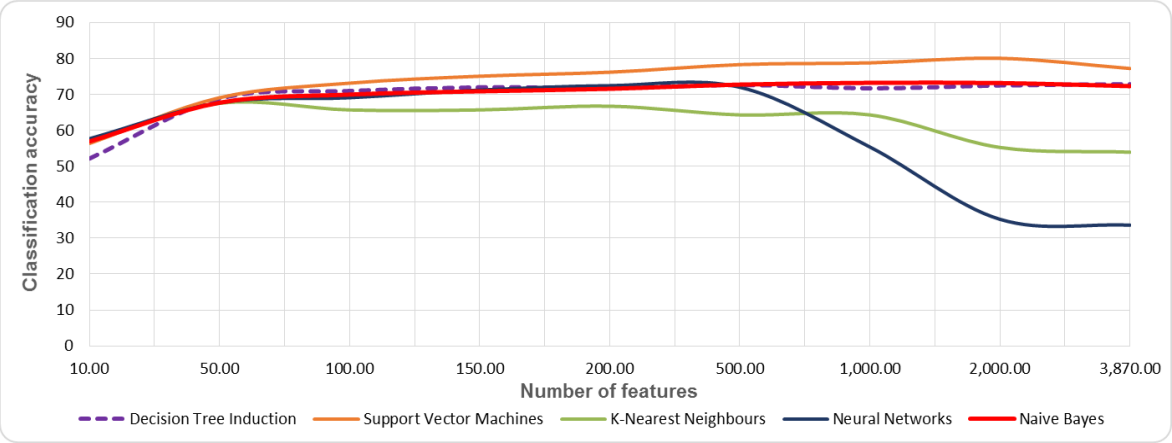

Figure 4.2 Influence of the feature size on classification accuracy for different supervised learning algorithms

Table 4.3 Accuracy of classification of individual web page types with decision tree induction.

Class	**Precision**	**Recall**	**F Measure**
About	0.59	0.46	0.52
Business and Innovation	0	0	0
Discussion	0.87	0.89	0.88
Research	0.63	0.8	0.7
Staff	0.78	0.75	0.77
Student Life	0.63	0.29	0.4
Study	1	0.33	0.5
Support	0.78	0.71	0.74

Table 4.4 Accuracy of classification of individual web page types with support vector machines.

Class	Precision	Recall	F Measure
About	0.59	0.62	0.60
Business and Innovation	0.00	0.00	0.00
Discussion	0.83	0.92	0.88
Research	0.68	0.81	0.74
Staff	0.82	0.74	0.78
Student Life	0.89	0.47	0.62
Study	1.00	0.17	0.29
Support	0.84	0.75	0.79

Table 4.5 Accuracy of classification of individual web page types with k nearest neighbours

Class	Precision	Recall	F Measure
About	0.36	0.62	0.46
Business and Innovation	0.00	0.00	0.00
Discussion	1.00	0.63	0.77
Research	0.64	0.71	0.64
Staff	0.79	0.73	0.76
Student Life	0.43	0.43	0.43
Study	0.00	0.00	0.00
Support	0.71	0.63	0.77

Table 4.6 Accuracy of classification of individual web page types with Naïve Bayes.

Class	Precision	Recall	F Measure
About	0.33	0.62	0.43
Business and Innovation	0.17	0.25	0.2
Discussion	1.00	0.68	0.81
Research	0.82	0.53	0.65
Staff	0.84	0.7	0.76
Student Life	0.43	0.43	0.43
Study	0.18	0.4	0.25
Support	0.61	0.78	0.68

Table 4.7 Accuracy of classification of individual web page types with a 3 layered neural network.

Class	Precision	Recall	F Measure
About	0.27	0.92	0.41
Business and Innovation	0.00	0.00	0.00
Discussion	0.91	0.78	0.84
Research	0.84	0.45	0.58
Staff	0.83	0.68	0.75
Student Life	0.00	0.00	0.00
Study	0.00	0.00	0.00
Support	0.86	0.7	0.77

The best result was from the SVM classifier, with an accuracy of 78%. It is also important to know how accurately the classifier identifies individual page types. This is determined using precision, recall and F measure, with results shown from **Table 4.3** to **Table 4.7**. Here, precision is the likelihood that the classifier will correctly classify a web page of type X as class X, while recall is the likelihood that the classifier will not classify a web page that is not of type X as class X. F measure is the accuracy of an individual class computed by the formula in Section 3.6.1 that depends on the precision and recall.

Table 4.8 Confusion matrix for classifying 2,549 web pages with support vector machines								
	About	**Business and innovation**	**Student life**	**Research**	**Staff**	**Study**	**Support**	**Discussion**
About	157	0	3	64	5	0	18	0
Business and innovation	3	0	0	7	0	0	2	0
Student life	21	0	16	2	0	0	6	0
Research	30	0	1	609	21	0	59	9
Staff	11	0	0	68	272	0	5	2
Study	2	0	0	13	1	8	4	1
Support	38	0	1	127	15	0	700	9
Discussion	3	0	0	17	0	0	5	214

Table 4.8 contains the confusion matrix from automatic classification of web pages with support vector machines. From **Table 4.8**, it suggests that the majority of wrongly classified Student Life web pages are misclassified as About web pages, while the majority of Business and Innovation web pages (web pages about university relationships with non-academic organisations) are misclassified as Research web pages. Hence, it is worth considering if certain web page types can be merged for a more appropriate classification scheme.

Figure 4.3 shows an example of a decision tree for the classification of the web pages. This is not the optimal result that can be achieved. The settings of the classifier were adjusted to reduce the size of the tree for

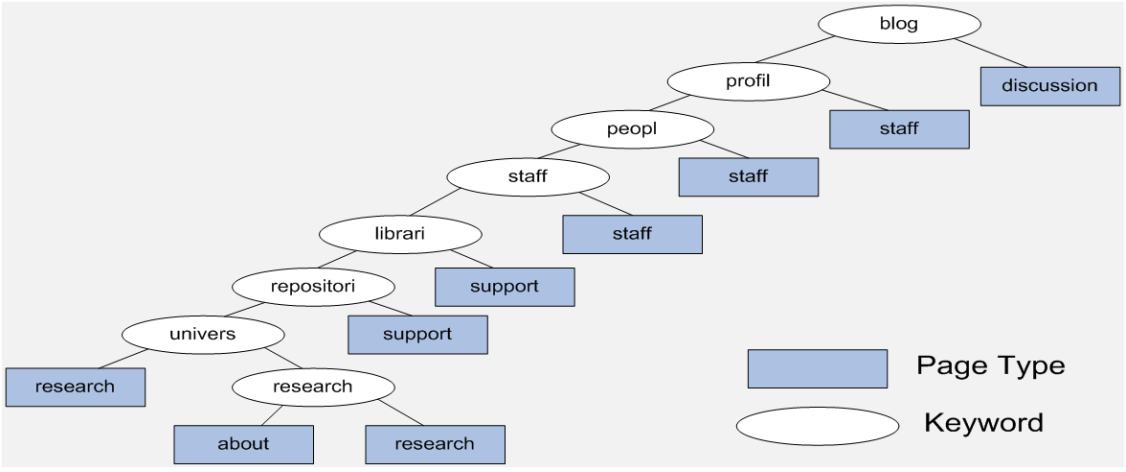

Figure 4.3 An example of a decision tree for the classification of web pages in universities' websites.

illustrative purposes. If the decision tree in **Figure 4.3** is used to classify university web pages, a number page types will always be incorrectly classified. Business and Innovation, Study and Student Life pages do not appear in any leaf node and so they will always be

misclassified. Web pages that do not contain any of the keywords will be classified as research pages. The classification model that constructed the tree in **Figure 4.3** had an accuracy of 46.8%.The nodes in the **Figure 4.3** show top terms in the feature set that are associated with a specific page type.

The accuracy of a supervised learning technique determines if it will be possible to use it in large scale studies. As the reported agreement between multiple coders in the manual classification of links for webometric research is between 70% and 98% (Holmberg, 2009), automatic classification of at least 70% may be acceptable for some types of webometric research. In this thesis (see **Table 4.2**), support vector machines automatically classified web pages with up to 78% accuracy. This accuracy may be acceptable because SVMs can classify hundreds of thousands of web pages significantly faster than manual approaches. However, the researcher has to decide the level of error acceptable by assessing the cost of misclassification compared to the gain of automation.

4.2.1. Predicting the target page type

If the reason for link creation in a university's website can be inferred from the relationship between the source and the target page, two web pages have to be visited by the software in order to extract features

that will determine their page types. In some cases, the target page may be unavailable, which makes using the relationship between web pages to suggest reasons for link creation impossible. If the target page type can be determined with information in the source page type, the classification process will be more efficient. Using information from the source page type can also ensure that the resulting category is closer to what the link creator in the source page thinks about the target page type.

Inferring the target page type can also be used for efficient crawler design, because page categories that are least likely to contain link data that helps reach the research goals will not be added to the frontier. Possible web page features that can be used to predict the target page type are:

- Link (URL) Anchor Hypertext
- Text around the link
- HTML tags – Thelwall (2003) suggests that links created because of collaboration are sometimes wrapped in an image HTML tag that links to the homepage of the collaborators.
- Source page type

The training data used is the same as the data used in Section 4.1.2. However, this training dataset (1,178 instances) is less than the dataset used in Section 4.1.2 because the training dataset in Section

4.1.2 is a combination of both the source and target web pages. Here, only the source web pages are used to predict the target web page.

Features used to predict the target web page type are the URL, anchor hypertext and text surrounding the link (four tokens to the left and right of the link). These features were represented as strings and pre-processed using the techniques described in Section 4.1.2.

Table 4.9 A comparison of the accuracy of 10 pre-processing options for decision tree induction and support vector machines, k nearest neighbours, Naïve Bayes and a 3-layered neural network supervised learning classifiers for predicting the target page type of 1,178 manually classified university web pages with baseline accuracy of 37.9%.

Bigrams /Unigrams	**TFIDF used rather than binary**	**Stemming**	**Stop words removed**	**Hypertext/ Text around included**	**URL included**	**DT Accuracy**	**SVM Accuracy**	**KNN Accuracy**	**Naïve Bayes Accuracy**	**MLP Accuracy**
*Bigrams + Unigrams	Yes	Yes	Yes	Yes	Yes	69.30%	73.90%	69.50%	61.80%	64.10%
Bigrams + Unigrams		Yes	Yes	Yes	Yes	68.00%	73.10%	68.10%	55.00%	65.50%
Unigrams	Yes			Yes	Yes	69.60%	72.70%	61.00%	64.40%	65.60%
Bigrams + Unigrams			Yes	Yes	Yes	69.80%	72.20%	67.80%	54.00%	64.40%
Unigrams	Yes	Yes		Yes	Yes	67.50%	70.20%	60.40%	64.60%	64.60%
Unigrams	Yes	Yes	Yes	Yes	Yes	66.10%	68.60%	58.80%	66.30%	60.30%
Unigrams		Yes		Yes	Yes	68.00%	68.00%	54.60%	59.80%	57.50%
Unigrams				Yes	Yes	71.30%	67.00%	56.50%	56.00%	58.30%
Unigrams		Yes	Yes	Yes	Yes	66.40%	65.50%	56.00%	57.20%	57.40%
Unigrams			Yes	Yes	Yes	67.80%	61.70%	56.10%	53.40%	55.50%

*Best pre-processing options and machine learning algorithm

The results in **Table 4.9** suggest that information in the source page can be used to determine the target page type with almost the same accuracy as using the web page URL and web page title of the target page.

4.2.2. Characteristics of outlinks in each web page category

Each page type was studied to identify the type of pages that they link to and possible reasons why the links were created. Reasons behind link creation that are specific to different web page types can be associated with individual links that belong to the page category.

The supervised learning techniques used in this thesis produced the best results when the web page title and URL determined the features of a web page used for automatic classification. The web page titles for the 1,824,582 unique web pages belonging to the UK university domains that the custom web crawler in Section 4.1 visited were not available. Only 5% of the web pages (97,299) were revisited to retrieve their page titles, in order to avoid re-crawling websites. The web pages were then automatically classified with SVMs using the methods described in Section 4.1. These automatically classified web pages were then manually analysed to identify the characteristics of outlinks in the different web page types. It should be noted that each of

the each of the 97,299 web pages were not rigorously analysed, results of analysis was based on a cursory overview.

Table 4.10 Web pages belonging to each page type from the 97,299 automatically classified university web pages and reasons link creation in different web page categories.

Source page type	No of web pages	Comments on links from these pages
Support	36,465	Rarely link to other page types Links to research pages that own or created the resource in the support page. Links are created to direct users to other relevant information, often to other pages that are created to improve learning, research or teaching skills.
Research	32,204	Links to About pages, usually a clickable logo of a collaborating university; organisational links as described by Thelwall (2003). Pages about research projects had links to staff pages of its collaborators or homepages of research groups or department. Research pages had links to all research groups or departments in the same scientific field.
Staff	9,690	Links to homepages of universities were often what Thelwall (2003) refers to as gratuitous links. Links to support pages that contain a resource, for example, staff publications. Links to other staff pages because of collaboration in a research project or co-authorship in a publication.
About	14,194	Rarely link to other universities. The majority of outlinks are for non-scholarly reasons.
Discussion	4,638	Links are created for a variety of reasons, so it is very difficult to identify a general pattern. Each blog entry belongs to a particular page type, and reasons for linking are the same as reasons of its corresponding page type.
Student Life (SL)	60	Mainly link to SL pages in close geographic locations. Perhaps this is a part of the reason why link analysis research shows that UK universities links to other universities in close geographic location.
Study	48	Most links are to support pages containing information relevant to the course. Links to research pages that contained software/ research output used in the course. Links to staff pages that authored course material, or a visiting professor Represents only a small set of total pages, perhaps because teaching materials are located on a protected server, thus inaccessible through public web crawlers.
Business and innovation (BI)	-	Usually link to non-university websites; its non-academic partners which explains its small size. Links to About pages; homepages of institutions working

		together in a community project.

4.3. Conclusions

The objective of this chapter was to determine how reliably link source and target pages can be automatically classified. To reach this objective, 7 categories that university web pages could belong to were identified using the three missions of HEIs as a guide. The method used to identify web page types is less detailed than the approach used in previous studies, but is a practical solution that makes automating the classification scheme easier. The manual classification of the training set that was used to create the classification model was done by a single researcher. This makes makes classification results subjective and undermines the reliability of the training dataset. Researchers can also create a different classification schemes depending on their research goals. If a public data set that is classified by multiple coders with an appropriate classification scheme agreed upon by the webometrics research community would be created, then it would eliminate the subjective nature of future webometrics studies.

In terms of the accuracy of determining the web page type using machine learning techinques, SVMs achieved up to 78% accuracy using the top 500 features with the highest information gain, derived from the web page URL and the web page title. Information from the source

page, that is, text around a URL and the URL itself can also be used to predict the target page type with up to 74% accuracy, which makes this method also applicable in the design of more efficient web crawlers because web page types that are least likely to contain the type of information required for a particular webometrics study could be blocked from the crawler frontier. The automatic classifier, however, always misclassifies business and innovation web page types. This is because of the small amout of business and innovation web sites in the training set. This suggests that in the classification scheme used in this study, web pages that belong to the business and information category are rare in a university's website. Perhaps a different classification scheme may be more suitable for automatic classification of university web pages.

Reasons for link creation specific to different web page types can be associated with individual links that belong to that web page category. Outlinks in different web page categories show common characteristics, and even outlinks to non-university websites sometimes point to domains that help to achieve the goal of a particular web page type. Support pages had links to the resource creator or teaching or learning pages. Staff web pages about a research project are ideal to identify collaboration or cooperation between institutions. Web pages about the living experience in the university, even if they represent only a small part, tend to link to web pages in close geographical regions.

The majority of these links are for non-scholarly reasons and should be excluded when identifying academic relationships between universities. Business and innovation and study web pages contributes less than 1% to the total links to other universities, perhaps because they are often inaccessible to web crawlers or the majority of their outlinks are to non-academic organisations. Administrative (about) pages also contained few links to other universities, and their outlinks were created for non-scholarly reasons. However, the majority of links to administrative (about) pages were either gratuitous or as a result of collaboration.

As different web page types show specific reasons for link creation, it shows that there is a possibility for using the relationship between the source and target page to analyse the reasons for linking in academic websites.

Finally, links between two staff web pages suggest collaboration between them and supervised learning methods can automatically identify staff web pages with high precision, 82%. Thus, there is a possibility for more in depth analyses of inter-linking between universities for collaboration studies, and staff links seem to be particularly important for this. The next chapter investigates the extent to which these links from university staff web pages indicate collaborative activities between universities.

5. Web data as an indicator for inter-university collaboration

This chapter addresses the second aim of the thesis; which is identifying a method that can be used to improve the quality of web based data, to make it more suitable for collaboration analysis between UK academic institutions. The result from two studies are presented in this chapter.

The extent to which hyperlink data can be improved as an indicator for collaboration is investigated in Section 5.1.

The suitability of web mentions as a collaboration indicator is investigated in Section 5.2. **Table 5.6** presents the results from a manual investigation of web mentions data, to identify if the reasons for web mentions in academic web sites is similar to the reasons for link creation. If they are similar, then the same machine learning techniques that were used in Chapter 4 for hyperlink analysis can also be used for web mentions.

The two studies carried out in this chapter is important not only for people who cannot afford bibliographic databases but also because webometrics indicators may be used to show collaborations that may not be reflected in traditional bibliometrics (Stuart, 2008).

5.1. Hyperlink data as an indicator of inter-university collaboration.

In what is perhaps the published literature most related to the study in this chapter, Stuart, Thelwall and Harries (2007) investigated if web links from university websites reflect collaborative relationships between the two organisations that the link connects. Their result suggests that direct linking cannot be confidently used to infer relationships between the two organisations that the link connects, but a significant proportion of outlinks from UK University websites reflect collaboration, so web links have the potential to be used as an indicator for collaboration, if methods for filtering out irrelevant hyperlinks are identified.

It is possible to automatically filter out some irrelevant hyperlinks. In Chapter 4 supervised learning techniques were used to automatically classify web pages in UK university websites. Library, information and career service pages, which were called support pages in the classification scheme showed high automatic classification accuracy, with a precision of up to 84%. Results in Chapter 4 show that this page type contributes up to 35% to the total web pages in a UK university's website and is unlikely to contain links that suggest collaboration (Stuart, Thelwall and Harries, 2007). Hence using supervised learning to automatically exclude such pages should help when using web links

as a collaboration indicator. Also, in chapter 4, a manual investigation of a random sample of links showed that the majority of the links between web pages related to staff in UK universities are created as a result of previous or future collaborations.

This study fills a gap in previous research because although hyperlinks have previously been used as indicators of academic collaboration (Stuart, Thelwall and Harries, 2007; Kretschmer, Kretschmer and Kretschmer, 2007), this has not been done in conjunction with automatic or manual web page classification. Academic links in this study are:

- Links between two universities' staff web pages or
- Links between university websites excluding those from library and other web pages created to provide service to university staff or students.

The goal of this study is achieved by answering the research questions:

(1) Can the extent of collaboration between two universities be better estimated with hyperlinks if only those links between university staff web pages are used rather than all links?

(2) Can the extent to which a university collaborates with other UK universities be better estimated by the total number of academic in-links rather than the total in-links to the university's website?

5.1.1. Methods

To determine if automatically identifying and restricting hyperlinks in university websites to only those that may have been created for collaborative reasons can help produce better collaboration indicators, web pages related to university staff and web pages that provide services to university staff or students were automatically identified using machine learning methods (support vector machines) because SVMs achieved the best results in chapter 4, and then statistical correlation tests were used to investigate if:

(1) The correlation between the number of links connecting two university websites and extent to which the two universities collaborate together is higher when only links connecting staff web pages are used than when all links connecting the two universities' websites are used.

(2) The correlation between the number of in-links to a university's website and the extent to which the university collaborates with all other universities is higher when only academic in-links are used rather than all in-links.

Co-authored publications have been widely used for collaboration studies (Ponds, van Oort and Frenken, 2007; Hoekman, Frenken and Tijssen, 2010; Hoekman, Frenken and Oort, 2008). Collaboration in

research projects have also been successfully used to study collaborative activities among organisations (Autant-Bernard et al., 2007; Ortega and Aguillo, 2010b), so in this study, the extent to which two universities collaborate together is estimated by the number of publications that the universities co-authored and the number of research projects that they co-participated in.

Co-authored publication data was only available for the 36 UK universities that appeared in the 2013 CWTS Leiden ranking, so even though it may be able to suggest how UK universities collaborate together, it is not sufficient to estimate the extent a UK universities collaborate with all other UK universities, so the extent to which a university collaborates with other universities is determined by the total number of research projects the university partook in with other universities, because research project participation data was available all UK universities.

5.1.1.1 Publication data

The number of co-authored publications between universities was extracted from publication data retrieved from the Web of Science. The Centre for Science and Technology Studies (CWTS) in Leiden University provided the processed co-authorship information for the 36 UK universities that appeared in their CWTS 2013 Leiden Ranking. Publication data is restricted to these 36 universities. These 36

universities had a total of 323,763 publications between 2008 and 2011.

The Leiden Ranking is based on publication data retrieved from the Thomson Reuters Web of Science bibliographic database. It ranks world universities based on their scientific impact, measured through citations and the extent in which they collaborate, measured through co-authorships. The data collection methodology used in the Leiden Ranking assigns publications to universities based on the institutional affiliation that authors indicate in their publication. The Leiden Ranking data collection has two stages, the first stage assigns a publication to a university when the university's address or variants of the university's address is explicitly mentioned, the second stage assigns publications of hospitals affiliated to a university to that university.

5.1.1.2 Project data

The UK research council's website (Research Council UK, 2013), Gateway to research (http://gtr.rcuk.ac.uk/), contains information about funded research projects from all UK research councils. A program was written to extract research project information for the 104 UK universities in Appendix A from the UK research council's website (Research Council UK, 2013) on 20th May, 2013.

To cover additional UK research funding, data from CORDIS (COmmunity Research and Development Information Service), a major source for EU research funding data was added to the data from UK research councils. Information about EU Funded research projects that started after 1 January 2006 was extracted from the CORDIS website. Records with incomplete data were excluded, so that 7,415 EU Funded projects and 30,091 UK research council funded projects were used for the analysis. These projects do not cover all UK universities' research funding between 2006 and 2013, but probably cover a large amount. In 2010/2011 the UK BIS Research Councils, the Royal Society and the British Academy were responsible for 35% of the UK Higher Education (HE) sector research income and European Union government bodies were responsible for 10% of the UK HE sector research income (Shef.ac.uk, 2014).

5.1.1.3 Hyperlink data

A custom web crawler was used to extract the links from one UK university to another. The crawler did not visit all webpages, it only covered the links that can be reached by iteratively following links from the university's homepage The pseudocode for the web crawler is described in Section 4.1. The dataset had 409,858 unique web pages from the universities in the 2013 Leiden Ranking.

5.1.2. Automatic web page classification

The dataset used in Section 4.1.2 was manually classified by the author of this thesis into staff pages and support pages as two separate facets. These manually classified pages formed the training set that was used to create the model for automatic web page classification.

Staff related web page facet:

- Staff pages: Related to staff in university. Examples include homepages, staff profiles, list of publications, CVs.
- Other pages: All other pages.

Academic pages facet:

- Support pages: Library web pages that contain repositories of learning resources for staff/student and documents/information for enhancing teaching/learning skills.
- Academic pages: All other pages.

The web pages in the training set were pre-processed and used to create the classification model. The following steps were used to transform the web pages into vectors for machine learning. The machine learning raw data was the web page URLs and web page titles

split into "tokens" at non alpha-numeric characters. All terms were converted to lower case and stemmed with the Porter Stemming Algorithm.

WEKA (Hall et al., 2009) implements support vector machines in its sequential minimal optimization (SMO) algorithm. The default settings of the classifier along with the top 250 features was used to create the model that automatically classified web pages. Results shown in **Figure 4.2** showed that the accuracy of support vector machines for this dataset is not greatly improved when more than 250 features are used. Tweaking the classifier settings or the number of features may improve the accuracy of the classifier but this was not necessary because the default settings produced up to 94.5% accuracy determined by a 10 fold cross validation on a human classified dataset.

When non academic web pages were excluded, the dataset contained 304,089 webpages. So approximately 25.8% of the web pages in the dataset was excluded.

5.1.3. Normalization

The size of a university is a factor that can influence the total number of inlinks and outlinks from a university's website. Academic staff size and research quality of universities are factors that influence the number of outlinks created in a university's website and the likelihood that it will

be the target of inlinks (Thelwall, 2002a). University size can also influence its total number of collaborations and therefore both collaboration and hyperlink data should be size normalized before conducting any correlation test.

The number of publications, the number of research projects, the number of inlinks and the number of project collaborations was divided by the total number of academic staff in 2008 for each university. The number of a university's project collaborations is the number of research projects the university partook in with other UK universities.

$$Inlinks\ per\ academic\ (IPA) = \frac{Number\ of\ inlinks}{Number\ of\ academic\ staff}$$

$$Academic\ inlinks\ per\ academic\ (\mathrm{A}IPA) = \frac{Number\ of\ academic\ inlinks}{Number\ of\ academic\ staff}$$

$$Publications\ per\ academic\ (PPA) = \frac{Number\ of\ publications}{Number\ of\ academic\ staff}$$

$$Project\ collaborations\ per\ academic\ (PCPA) = \frac{Number\ of\ research\ project\ collaborations}{Number\ of\ academic\ staff}$$

$$Research\ projects\ per\ academic\ (RPPA) = \frac{Number\ of\ research\ projects}{Number\ of\ academic\ staff}$$

The number of links connecting two universities, the number of co-authored publications and the number of co-participating projects

between two universities was divided by the product of number of academic staff in the two universities in 2008 for normalization.

$$\text{Normalized number of coParticipating projects } (NCP) = \frac{\text{Number of projects universites } x \text{ and } y \text{ coparticipated in}}{\text{No of academic staff in } x * \text{No of academic staff in } y}$$

$$\text{Normalized number of links } (NL) \text{ between universites } x \text{ and } y = \frac{\text{Number of links from } x \text{ to } y + \text{number of links from } y \text{ to } x}{2 * \text{No of academic staff in } x * \text{No of academic staff in } y}$$

$$\text{Normalized staff target links (NSTL) between universites } x \text{ and } y = \frac{\text{Number of links from } x \text{ to a staff page in } y + \text{Number of links from } y \text{ and to a staff page in } x}{2 * \text{No of academic staff in } x * \text{No of academic staff in } y}$$

$$\text{Normalized number of inter staff links } (NISL) \text{ between universites } x \text{ and } y = \frac{\text{No of links from a staff page in x to a staff page in yo of links from a staff page in y to a staff page in x}}{2 * \text{No of academic staff in } x * \text{No of academic staff in } y}$$

$$\text{Normalized number of coAuthored publications } (NCAP) = \frac{\text{Number of coAuthored publications}}{\text{No of academic staff in } x * \text{No of academic staff in } y}$$

The Higher Education Statistics Agency (www.hesa.ac.uk) was used to provide data about the number academic of staff in each UK university.

The geographic distance separating two universities was determined by the straight line distance between the addresses of the two universities. Longitude and latitude geographic coordinates of the addresses of universities was extracted from Google Maps. The spherical law of cosines was used to compute the geographic distance separating two universities. This formula is widely used to compute the distance between two geographic coordinates (Hasan, Rahman and Haque, 2009).

$$Geographic\ distance\ separating\ universites\ x\ and\ y\ in\ km = R * (\arccos(\sin(latX) * \sin(latY) + \cos(latX) * \cos(latY) * cos(longY - longX)))$$

R is the radius of the earth; 6,371km. LatX and LongX are the latitude and longitude geographical coordinates of university X, while LatY and LongY are the latitude and longitude geographical coordinates of university Y. Longitude and latitutude coordinates from Google maps were converted from degrees to radians.

5.1.4. Results

To identify the extent of association between hyperlinks and publication data from Thomson Reuters, the data was compared using Spearman correlations. **Table 5.1** and **Table 5.2** show the correlations between

hyperlink data, research project data and publication data of UK universities.

Table 5.1 Spearman correlations between the links between two universities' websites (NL), the staff target links (NSTL), inter-staff links (NISL), the number of co-participating projects (NCPP), co-authored publications (NCAP) and the geographic distance separating two UK universities in the 2013 CWTS Leiden Ranking (all normalized except distance).

	NL	**NSTL**	**NISL**	**NCPP**	**NCAP**	**DISTANCE**
NL	1	0.614**	0.431**	0.271**	0.381**	-0.461**
NSTL		1	0.687**	0.316**	0.412**	-0.447**
NISL			1	0.259**	0.377**	-0.360**
NCPP				1	-0.284**	-0.334**
NCAP					1	-0.367**
DISTANCE						1

** Correlation is significant at the 0.01 level (2-tailed).

Table 5.2 Spearman correlations between the total number of inlinks (IPA), academic inlinks (AIPA), publications (PPA), project collaborations per academic (PCPA) and research projects (RPPA) for UK universities (all per academic).

	IPA	**AIPA**	**PPA**	**PCPA**	**RPPA**
IPA	1	0.784**	0.181	0.570**	0.586**
AIPA		1	0.213	0.750**	0.774**
PPA			1	0.421*	0.483**
PCPA				1	0.975**
RPPA					1

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Statistical correlations between links and collaboration indicators may be a result secondary factors that influence both links and co-authored publications. Webometrics results can be only partially validated through correlation tests (Thelwall, 2004), which is why a random selection of links should be also classified to give context to the results of webometrics studies (Thelwall, 2009, 2006). However, if links are previously selected to include only those links that may be created for collaborative reasons before the correlation tests, it can increase the confidence about the validity of correlation tests.

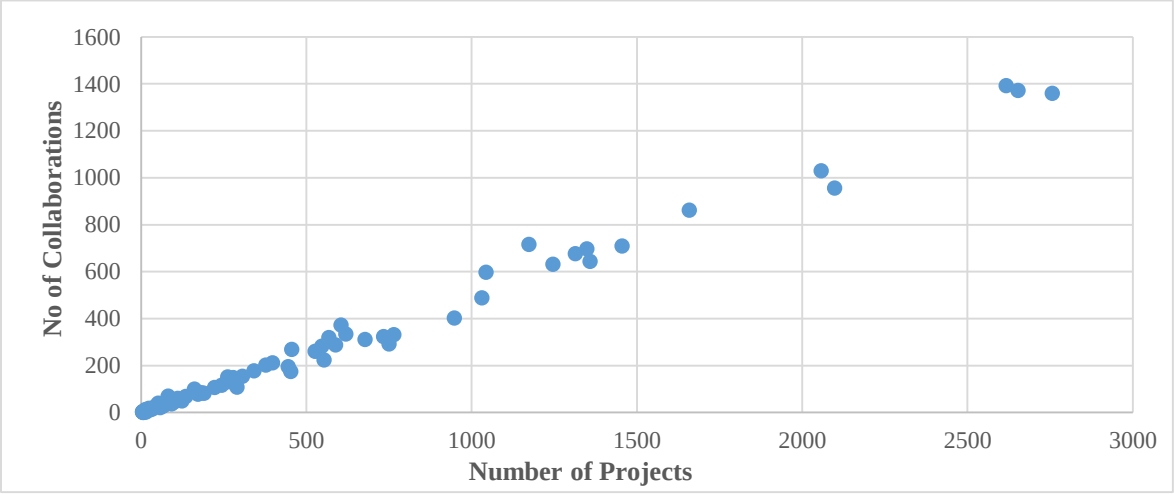

Figure 5.1 Linear relationship between the number of research projects an institution participated in and the number of research projects the university partook in, in collaboration with another institution.

Correlation results in **Table 5.2** show that there is significant increase in the correlation between collaboration in terms of co-

participation in research projects and links when only academic links are used as instead of all links. This suggests that machine learning techniques can improve to quality of hyperlink data for inter-university collaboration studies. There is also marginal increase in the correlation between links and the number of co-authored publications when hyperlink data is pre-processed with supervised learning techniques.

Collaboration has been shown to positively affect productivity in terms of the number of publications (Lee and Bozeman, 2005; Landry, Traore and Godin, 1996; Katz and Martin, 1997; Subramanyam, 1983), the high correlation in **Table 5.2** and linear relationship in **Figure 5.1** suggests that collaboration also positively relate to the number of research projects a university carries out.

5.1.5. Discussion and Conclusions

This study investigated the extent to which hyperlink data can be improved to adequately indicate the extent of collaboration between universities. Hyperlinks connecting UK University websites were analysed to find if there is any association between hyperlink data and traditional data (co-authored publications from databases like Thomson Reuters or Scopus) that are used to study collaboration between organisations.

The first research question investigated if the extent of collaboration between two universities will be better estimated by links if only links between staff related web pages are used. When university staff-related web pages were automatically identified using a supervised learning technique with up to 94% accuracy, the correlation between the number of links between university staff related web pages and the co-authored publications between the two universities was lower than the correlation between all the links between the two university websites and co-authored publications. Even though the majority of links between staff related web pages may have been created for collaborative reasons, inter-staff links do not associate better with other indicators of collaboration compared to all links. This may be because, currently, there are not enough staff pages in university websites to produce more accurate results. Also, as a result of the changing web, some academics may have their online presence in social networking sites like LinkedIn, Facebook, Twitter and Acedemia.edu instead of personal web pages and online CVs. It can be time cosuming to find user profiles on social media websites (Holmberg and Thelwall, 2014), but methods that incoporate data from social media websites with other web sources should be developed for future webometrics research. Although most researchers still maintain the main characteristics of research dissemination, gradually more researchers are also using social media to dessiminate their research as they understand the potential these social media tools bring (Weller, 2011).

Academic institutions also have social media presence. However most universities use their social media accounts for media and communication. These accounts are similar to the "About web page category" in the classification scheme in **Table 4.1**.

Links between staff pages are not the only links in an institution's website that suggest collaboration. Previous studies show that links that appear at shallow depths and links that are clickable logos are likely to indicate collaboration relationships between the two organisations that the link connects. Including links from other webpages that contain links created for collaboration reasons may therefore improve the results. There is marginal improvement in the results when links are restricted to staff target web pages instead of inter-staff links or raw links.

5.2. University name mentions as an indicator for inter-university collaboration.

The second study is this chapter investigates if the web mentions in universities' websites can be used as an alternative to hyperlinks to

study the extent to which two universities collaborate together. The following research questions help to reach the objective of this study:

1) Does the number of mentions of a university's name in another university's website correlate with the traditional indicators of collaboration between two universities?

2) Is the correlation between university name mentions and traditional indicators of collaboration higher than the correlation between hyperlink counts and traditional indicators of collaboration?

3) What are the reasons for university name mentions in the different web page categories in a typical university's website?

In this study, traditional indicators for collaboration between two universities are the number of co-authored publications between the two universities and the number of research projects that the two universities co-participated in, because both co-authored publications (Hoekman, Frenken and Tijssen, 2010) and co-participation in research projects (Autant-Bernard et al., 2007) have been widely used for collaboration studies.

5.2.1. Methods

Correlation tests were used to investigate the association between university name mentions and the traditional indicators of collaboration.

Correlation tests were also used to determine if university name mentions associate better with other collaboration indicators than hyperlink counts do.

University name mentions were compared with collaboration in terms of co-authored publications and co-participation in research projects to assess through correlation the degree of association between university title mentions and collaboration in terms of co-authored publications, and university title mentions and collaboration in terms of co-participation in research projects.

Because significant correlation between variables may be because of other common factors present in the two variables, a content analysis was also performed to determine the extent to which a university's name mentioned in another university's website was because of collaboration between the two universities.

5.2.1.1 Data

Publication data is the same as data in Section 5.1.1.1, research project data is the same as data in Section 5.1.1.2, and hyperlink counts between universities' websites is the same as data in Section 5.1.1.3 . Hyperlink counts, the number of co-participating projects and the number of co-authored publications were size normalized using techniques described in Section 5.1.3.

The number of mentions of a university in another university's website was determined by the Hit Count Estimate (HCE) of a query passed to the Google search engine. For example, the query ("oxford university" | "university of oxford" site:warwick.ac.uk) was queried to get an estimate of the number of times the University of Oxford was mentioned in Warwick University's website. A number of universities have multiple domain names, so the HCEs of queries for each of the university's domain names were summed to get the final estimate of university mentions.

The number of university name mentions was also normalized by the number of academic staff in the two universities.

$$Normalised\ number\ of\ times\ university\ x\ mentions\ university\ y\ \text{(NM)}\ in\ its\ website = \frac{Number\ of\ mentions\ of\ y\ in\ x}{No\ of\ academic\ staff\ in\ x * No\ of\ academic\ staff\ in\ y}$$

5.2.1.2 Content analysis

To determine possible reasons for university name mentions in the web page types listed in **Table 4.1** and to investigate the extent to which university name mentions are because of collaboration, search results from the Bing search engine were downloaded using Webometric Analyst (Thelwall and Sud, 2012), automatically classified using the supervised learning techniques described in Section 4.1.2 and manually

studied to determine reasons for mentions in the different web page types.

Bing is one of the few remaining search engines that allow applications to automatically download search results and it allows up to 5000 free queries every month. The maximum normal number of query matches using the Bing API is 1000 (Thelwall and Sud, 2011). Significantly more results can be retrieved through query splitting (Thelwall, 2008a), but at the cost of the monthly query allowance. In some cases, the HCE of university mentions in another university website can be up to 20,000.

The search queries passed to Google was queried to Bing through the Webometric Analyst application. The queries sent to Bing and Google for the University of York are in Appendix F. The university names in the queries were modified for the other 35 universities in the Leiden 2013 ranking to extract their web mentions from Google and Bing.

The search results retrieved with Webometric Analyst contain additional information about each search result: the URL, web page title and a short description of the result. This additional information can be used as features of a supervised learning algorithm that map web pages to output categories.

SVM was used to automatically classify web pages in to categories in **Table 4.1** and 50 random web pages from each web page type were then manually examined to determine possible reasons for university name mentions in the different web page types and to determine if university name mentions in certain page types are most or least likely because of collaborative reasons.

5.2.2. Results

Table 5.3 Distribution of 313,294 web pages retrieved from BING that were automatically classified with support vector machines into the different web page categories in **Table 4.1**.

Web page category	**Size**
About	13.37%
Discussion	2.68%
Research	46.20%
Staff	22.16%
Study	0.09%
Support	15.47%

Table 5.4 Spearman correlations between the number of university mentions (NM), links (NL), the number of co-participating projects (NCPP) and co-authored publications (NCAP) between two UK universities (all normalized by dividing by staff numbers).

	NM	**NL**	**NCPP**	**NCAP**
NM	1	0.385**	0.121*	0.318**
NL		1	0.271**	0.381**
NCPP			1	0.284**
NCAP				1

** Correlation is significant at the 0.01 level (2-tailed)

*Correlation is signification at the 0.05 level (2-tailed)

University name mentions had lower correlations with other indicators of collaboration compared to hyperlinks. This suggests that hyperlink counts may be better than web mentions as an indicator for investigating the extent two universities collaborate together. However, simple hyperlink counts and web mentions are still unreliable indicators as shown in the correlation results in **Table 5.4** and **Table 5.1**, although hyperlinks and name mentions may not correlate with bibliometrics indicators because bibliometrics indicators do not reflect all types of collaboration (Stuart, 2008).

The size of a university is a factor that significantly influences the number of links and collaborations between academic institutions. This is evident in the significant correlation between the un-normalized variables in **Table 5.5** and the low correlations between the size normalized variables in **Table 5.4**.

Table 5.5 Spearman correlations between the number of mentions between two universities, links between two universities, projects two universities co-participating in (CPP), publications two universities' co-authored (CAP) and the product of the academic staff (AS) in the two universities (not normalized).

	Mentions	**Links**	**CPP**	**CAP**	**AS**
Mentions	1	0.633**	0.550**	0.683**	0.650**
Links		1	0.534**	0.634**	0.493**
CPP			1	0.705**	0.698**
CAP				1	0.803**
AS					1

**Correlation is significant at the 0.01 level (2 tailed)

Table 5.6 Reasons for university name mentions in the different university web page categories

Page Type	**Some comments about possible reasons for mentions in web page category**
About	The two universities are working together to provide community or social service like protection of children or the environment. Profile of individuals given honorary degrees, professorships or new appointments. Profiles include university the individual graduated from or his/her affiliation. University press release about new research project collaboration or partnership with another university.
Discussion	Affiliation of speaker in seminar, lecture or events the blogger is writing about or attending. The profile of the blogger, his/her second affiliation or university s/he graduated from. Research blogs about ongoing research projects and its collaborators or credit to research output from another university. Mention in blog roll – other blogs the blogger likes.

	Blog owner's collaborator in another university.
Research	Affiliation of speaker, chair or participant in a conference, workshop or seminar. Publisher of research group's article e.g. Oxford press or location of conferences attended. Collaborators in workshop or research projects. Co-authors affiliation in a scientific publication. Mention in a rich file (e.g. pdf, doc)
Staff	Second affiliation, previous affiliation or institution the staff graduated from. Affiliation of staff collaborator. Publisher of staff article (Oxford press, Cambridge press) or location of conference staff attended. Acknowledgement in staff publication.
Study	Location of publication or material in course reading list University course teaching staff graduated from. Information about summer school speaker from another university. About 90% of web pages were from the University of Surrey; perhaps study web pages from other universities are not accessible because they are on a secure server.
Support	Mention in title of document. Profile of speaker in skill and development workshop. In meta data of library item – publisher, author affiliation or reference. Location document was gotten from.

5.2.3. Discussion and Conclusions

This study investigated the extent to which the number of web mentions of a university in another university's website can be used as an indicator to investigate inter-university collaboration.

The degree of association between web mentions and collaboration in terms of co-authored publications and co-participation in funded research projects was analysed through correlation. The

significant correlation between non-normalized web mention counts and traditional indicators of collaboration confirms from previous research that the size of an institution is a factor that influences the number of in-links to and out-links from university websites. The correlation between the normalized number of mentions and the other indicators of collaboration was low, however. Based on the correlation results, hyperlinks associate marginally more strongly than web mentions with collaboration.

Like links, the majority of mentions of another university seem to be for scholarly reasons. Reasons for hyperlink creation in different university web page types are broadly the same as the reasons for mentioning a university. However, in a number of cases where web mentions were created explicitly for collaboration, there were no visible links to the collaborating university. This highlights the need for a systematic combination of different web based data sources (e.g., co-word occurrences, URLs and web mentions) to more effectively use web based data to study collaboration between organisations. The next chapter investigates what can be extracted from co-word analysis of the text in computer science departments' homepages.

6. Cluster analysis of computer science research groups

The final empirical chapter addresses the third aim of the thesis; an exploratory analysis of research group homepages in order to find what could be identified through unsupervised learning (clustering).

Computer science research groups are clustered based on the text in their homepages to determine the similarities in the research interests of various computer science departments in the UK. The purpose is primarily to assess the value of machine learning in this context rather than to investigate computer science research in particular.

Subsequent sections describe the data collection methods and clustering analysis through self-organising maps and principal component analysis.

The conclusion describes the findings from the exploratory analyses, showing that these type of analyses can be useful for policy makers to identify suitable collaborators or identify groups that can be merged to foster inter-disciplinary collaboration.

6.1. Text analysis to identify related computer science departments

Word co-occurrences in the homepages of research groups to investigate the relatedness of university departments. The following research question is drives this study:

1) Can an unsupervised machine learning cluster analysis of the text in the homepages of computer science research groups in the UK with self-organising maps and principal component analysis reflect similarities in interests between the departments?

Science and Engineering web pages are dominant in academic institutions web sites (Thelwall and Price, 2003), particularly computer science related web pages, which are heavily represented in academic websites (Thelwall et al., 2003). As this study is primarily exploratory, the research groups are restricted to only those that are computer science related, because computer science dominates online web presence in university websites compared to other disciplines. In future studies, it can be extended to include other disciplines and thus analyse whole universities based on their research interests. The study is also restricted to one country, the UK, for practical reasons.

Subsets of computer science departments are research groups, which define the key research interests of that department. The homepage of research groups describes their current research areas and sometimes advertise for PhD opportunities. So, analysis of the text in these homepages may be able to give insights into the research interests of the research groups, and at a higher level the research interests of the departments.

In webometrics research, co-word analysis have been investigated as an alternative for co-link analysis and used to determine the relatedness of organisations (Vaughan and You, 2010). If co-word occurrences can be used to measure the relatedness of organisations, and more related organisations collaborate together more often than less related organisations (Thijs and Glänzel, 2010), then a cluster analysis of computer science research groups based on the text in their homepages may be able to identify computer science departments with shared research interests (determined by the extent to which their research groups appear in the same cluster) who are collaborating together or may benefit from collaboration.

SOM is used as the clustering algorithm because of its unique property; the resulting clustering solution is a map, where similar clusters are placed closer together whilst less similar clusters are placed farther apart. Other clustering techniques described in the literature

review (see 2.4.2 page 53) may also produce adequate clustering solutions, but they do not have the unique SOM property.

6.1.1. Methods

The research groups were identified by manually visiting the homepages of 76 computer science departments and then identifying their research groups from the links in the homepages. In several cases, the homepage of the research group was not in the university domain. For example, Heriot-Watt's Interaction Lab was located in the Google domain.

372 research group homepages (listed in **Appendix B)** were identified from the computer science departments' websites visited. On average, each department had 5 research groups, although 21 research groups were identified in Imperial College London, the next highest was 15 identified in the University of Edinburgh. Other universities like Oxford Brookes and Aston University did not have dedicated homepages for each research group; a single website described the research of the whole department.

A Java program was written to automatically visit these homepages, extract the text content of the homepages whilst stripping out scripts, code, meta data and HTML tags, and then represent the text as non-alphanumeric delimited n-grams (n=1, 2 and 3). Here, n =

3 is used as the maximum because, in the list of computer science key phrases published in Microsoft's (2012) website, only 2% of the top 5000 computer science key phrases contained more than 3 words. As of February 2012, when the computer science key phrases were downloaded, the list had 39,458 key phrases that had appeared in at least one computer science publication.

In order to minimise the influence of non-computing terms on the clustering solution, the n grams from each web page were pre-processed to include only n grams in the list of computer science keywords published by Microsoft.

$$Final\ set\ of\ tokens\ in\ research\ group\ i\ (X_i) = Inital\ set\ of\ n\ grams\ \cap computer\ science\ keywords$$

Each research group was then represented as a vector of term frequencies multiplied by their inverse document frequencies (tf-idf). There were 749 unique terms in the keyword vocabulary, so each research group vector had a length of 749.

$$Research\ Group\ i = [TFIDF_1, TFIDF_2, \ldots\ , TFIDF_{749}]$$

6.1.1.1 Clustering

Factor analysis and Singular Value Decomposition (SVD) can be used

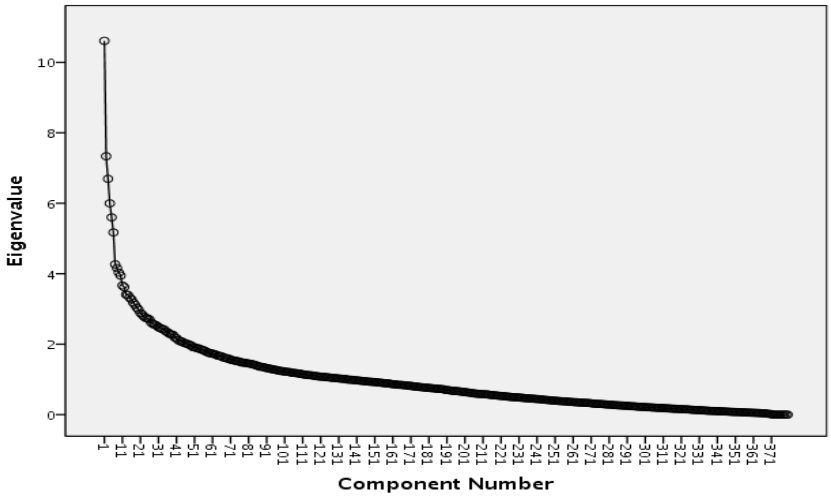

Figure 6.1 Scree plot of the PCA of tf-idf vectors of computer science key phrases

for clustering documents but SVD analysis is not easily available in popular statistical packages (Leydesdorff and Welbers, 2011), which is one reason why factor analysis may be preferred to SVD for clustering purposes. Ding and He (2004) have also shown that Principal Component Analysis (PCA) can be seen as solutions to the k-means objective function, so PCA may be used for feature extraction in machine learning problems.

A PCA was performed on the research groups' tf-idf vectors of computer science key phrases, and then the components with eigenvalues greater than one were extracted from the un-rotated factor

solution and used as input for the construction of SOMs. The scree plot from the PCA is in **Figure 6.1**.

SOM clustering results can be used to validate whether the PCA components are reliable features for the clustering problem in this study, and if principal components are valid clustering solutions.

6.1.1.1.1 Construction of Self-Organising Maps

This study used an implementation of SOM in Matlab (Vesanto et al., 1999) as previously used in other research (Singh et al., 2013; Hayfron-Acquah and Gyimah, 2014; Olawoyin et al., 2013).

The input parameters for the construction of SOMs can influence the quality of the clustering solution (Ballabio, Vasighi and Filzmoser, 2013). If it is difficult to find the optimal input parameters then exhaustive search can be used to identify them (Marini, Zupan and Magrì, 2004), although it can be time consuming.

The function to create a SOM in Vesanto et al.'s (1999) Matlab implementation is som_make. This function has a number of input parameters:

Algorithm: The algorithm of the SOM can either be batch or sequential. A sequential algorithm updates the neurons on the SOM by iteratively introducing a random sample from the input one at a time to while a batch algorithm introduces all samples to

the SOM at once, computes the winning neurons and then updates the weight of all neurons at the same time.

Initialization: Possible values of the initialization parameter are 'randinit' or 'lininit'. Randinit initializes the weight of each neuron on the SOM to a random value between the maximum and minimum values in the input, while lininit initializes the weights based on the eigenvalues of the input. When lininit initialization and a batch training algorithm is used, the final weights of the SOM is always the same (Ballabio, Vasighi and Filzmoser, 2013).

Lattice: The lattice of the SOM can be either hexagonal or rectangular. The difference between the two is largely in the number of neighbours each neuron can have, neurons in the hexagonal grid have more neighbours than neurons in the rectangular grid, as illustrated in **Figure 6.2**.

Neighbourhood function: The neighbourhood function is the mathematical function that is used to determine the extent to which neighbouring neurons are adjusted to the weight of the winning neuron or best matching unit (BMU). In the toolbox, the neighbourhood function can be 'gaussian', 'bubble', 'cutgauss' or 'ep'; based on different mathematical functions.

Size: The size of the map determines the number of neurons. For example a 20x20 map will have 400 units/neurons. In this study the size was set to be in the range [2, 20], 20 being when each research group can have its own neuron.

Training: The training parameter determines the number of epochs (times an input sample can be introduced to the map) and the learning rate for the SOM algorithm. The toolbox determines the most appropriate learning rates based on the size of the map and the specified training parameters which can be either 'short', 'long' or 'default'.

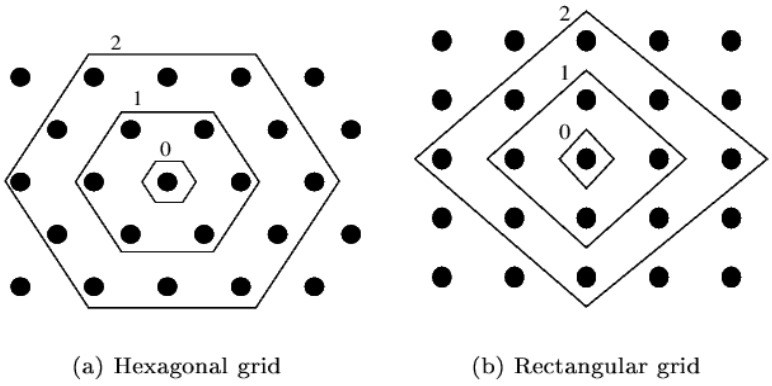

Figure 6.2 Neighbourhood size of the hexagonal and rectangular SOM lattice (Vesanto et al., 1999).

The quality of a resulting SOM can be analysed in terms of hits, topographical error (TE) or quantization error (QE). The number of hits is the number of times each neuron on the SOM was the best matching neuron, QE is the average Euclidean distance between each sample and its best matching neuron, and topographical error is the percentage of input samples where the best matching neuron and second best matching neurons are neighbouring neurons.

Quantization error reduces as the size of the SOM increases (Sun, 2000), and intuitively the number of hits will also increase with increase

in the map size (number of neurons), while topographical error will increase with increase in the size of the SOM. So, these quality metrics are best used to compare SOMs of relatively identical sizes. Ultimately, clustering solutions should also be manually verified to determine if the results make sense.

In multiple runs (1000), a possible value for each input parameter was randomly selected, and SOM constructed with these parameters. The quality of the resulting SOM was computed in terms of the hits, quantization error and topographical error. 1000 runs were heuristically determined to be able to achieve close to the optimal SOM clustering solution.

6.1.2. SOM clustering result

Table 6.1 to **Table 6.5** report the top 10 combinations of input parameters with the lowest quantization error. Not surprisingly, the quantization error was lowest for the largest sized (20) SOM, however the quantization error seemed to increase as the dimension of the input samples (principal components) increase. Using fewer components improves the time it takes to construct an SOM, and it is significantly faster than using the tf-idf vectors as input. It is however essential to also evaluate the topological ordering of SOMs in order to identify if the clustering solutions are adequate.

Figure 6.3 shows the topological ordering of SOMs when the principal components are used as the input. The research themes, based on the names of the research groups each cluster predominantly show research groups from either the same university or groups in similar research areas. Some clusters show research groups from the same university because some research groups' homepages have links to all other computer science research groups from that university. The link hypertexts were descriptions of the target research group, and this text influences the final cluster that a research group belongs to.

Interestingly, different PCA components seemed to be responsible for particular clusters on the map. For example, the value of the 2nd component was high for those research groups in the natural language processing theme and the research groups in the software engineering theme had high values for the 3rd PCA component. This is also in line with (Leydesdorff and Welbers, 2011) that show evidence that factors of PCA can be used for semantic mapping of documents.

As particular PCA components can determine the cluster a sample will belong to, a simple clustering algorithm that groups research groups based on each component may be able to identify departments with identical research interests, depending on how often they appear in the same group.

Table 6.1 The number of hits, quantization error (QE) and topographical error (TE) of the top 10 combinations of the input parameters with the lowest QE, after 1000 runs of the SOM algorithm. The first two components of the PCA of the tf-idf matrix is the input data.

Index	Algorithm	Initialization	Lattice	NF	Training	Size	Hits	TE	QE
198	batch	lininit	rect	cutgauss	long	20	400	0.1711	0.0052
913	batch	randinit	rect	Ep	long	20	400	0.2447	0.0053
77	batch	lininit	rect	cutgauss	default	20	400	0.2105	0.0055
882	batch	lininit	rect	cutgauss	default	20	400	0.2105	0.0055
276	batch	lininit	hexa	cutgauss	long	20	400	0.0500	0.0055
510	batch	randinit	rect	bubble	long	20	400	0.1816	0.0056
74	batch	lininit	rect	bubble	short	20	400	0.2237	0.0057
813	batch	lininit	rect	bubble	short	20	400	0.2237	0.0057
280	batch	randinit	hexa	Ep	long	20	400	0.0658	0.0058
293	batch	randinit	hexa	gaussian	long	20	400	0.0816	0.0058

Table 6.2 The number of hits, quantization error (QE) and topographical error (TE) of the top 10 combinations of the input parameters with the lowest QE, after 1000 runs of the SOM algorithm. The first 10 components of the PCA of the tf-idf matrix is the input data.

Index	Algorithm	Initialization	Lattice	NF	Training	Size	Hits	TE	QE
326	batch	randinit	rect	gaussian	long	20	400	0.0816	0.0820
234	batch	randinit	rect	gaussian	default	20	400	0.1237	0.0822
73	batch	lininit	rect	gaussian	long	20	400	0.0816	0.0831
143	batch	lininit	rect	gaussian	long	20	400	0.0816	0.0831
232	batch	lininit	rect	cutgauss	long	20	400	0.0816	0.0831
83	batch	randinit	rect	cutgauss	long	20	400	0.1053	0.0832
426	batch	randinit	rect	cutgauss	long	20	400	0.1421	0.0835
759	batch	randinit	rect	bubble	long	20	400	0.1605	0.0835
419	batch	lininit	rect	gaussian	default	20	400	0.0974	0.0837
489	batch	lininit	rect	Ep	default	20	400	0.0974	0.0837

Table 6.3 The number of hits, quantization error (QE) and topographical error (TE) of the top 10 combinations of the input parameters with the lowest QE, after 1000 runs of the SOM algorithm. The first 20 components of the PCA of the tf-idf matrix is the input data.

Index	Algorithm	Initialization	Lattice	NF	Training	Size	Hits	TE	QE
274	batch	lininit	rect	gaussian	long	20	400	0.0605	0.1555
314	batch	lininit	rect	gaussian	long	20	400	0.0605	0.1555
538	batch	lininit	rect	cutgauss	long	20	400	0.0605	0.1555
582	batch	lininit	rect	gaussian	default	20	400	0.0921	0.1566
127	batch	randinit	rect	gaussian	short	20	400	0.0763	0.1591
790	batch	randinit	rect	cutgauss	short	20	400	0.0737	0.1592
471	batch	lininit	rect	bubble	short	20	400	0.1342	0.1610
597	batch	lininit	rect	cutgauss	short	20	400	0.1342	0.1610
641	batch	lininit	rect	bubble	short	20	400	0.1342	0.1610
978	batch	randinit	hexa	gaussian	long	20	400	0.0237	0.1612

Table 6.4 The number of hits, quantization error (QE) and topographical error (TE) of the top 10 combinations of the input parameters with the lowest QE, after 1000 runs of the SOM algorithm. All components of the PCA with eigenvalue more than one of the tf-idf matrix is the input data.

Index	Algorithm	Initialization	Lattice	NF	Training	Size	Hits	TE	QE
870	seq	lininit	rect	gaussian	long	20	400	0.0053	0.5652
196	seq	lininit	rect	bubble	long	20	400	0.0053	0.5668
124	batch	lininit	rect	bubble	default	20	400	0.0237	0.5700
173	batch	lininit	rect	Ep	default	20	400	0.0237	0.5700
529	batch	lininit	rect	cutgauss	default	20	400	0.0237	0.5700
195	batch	lininit	rect	Ep	long	20	400	0.0079	0.5700
827	batch	lininit	rect	gaussian	long	20	400	0.0079	0.5700
434	seq	lininit	rect	gaussian	long	20	400	0.0053	0.5703
631	batch	randinit	rect	cutgauss	long	20	400	0.0079	0.5705
967	batch	randinit	rect	bubble	long	20	400	0.0105	0.5713

Table 6.5 The number of hits, quantization error (QE) and topographical error (TE) of the top 10 combinations of the input parameters with the lowest QE, after 1000 runs of the SOM algorithm. The tf-idf matrix is the input data.

Index	Algorithm	Initialization	Lattice	NF	Training	Size	Hits	TE	QE
902	batch	randinit	rect	bubble	short	20	400	0.0395	0.9566
44	batch	lininit	rect	bubble	default	20	400	0.0237	0.9636
336	batch	lininit	rect	cutgauss	default	20	400	0.0237	0.9636
497	batch	lininit	rect	gaussian	default	20	400	0.0237	0.9636
116	batch	randinit	rect	bubble	default	20	400	0.0132	0.9645
368	batch	randinit	rect	Ep	short	20	400	0.0526	0.9663
843	seq	lininit	rect	bubble	long	20	400	0.0132	0.9665
788	batch	lininit	rect	Ep	long	20	400	0.0237	0.9671
364	batch	lininit	rect	cutgauss	short	20	400	0.0316	0.9678
98	seq	lininit	rect	gaussian	long	20	400	0.0158	0.9689

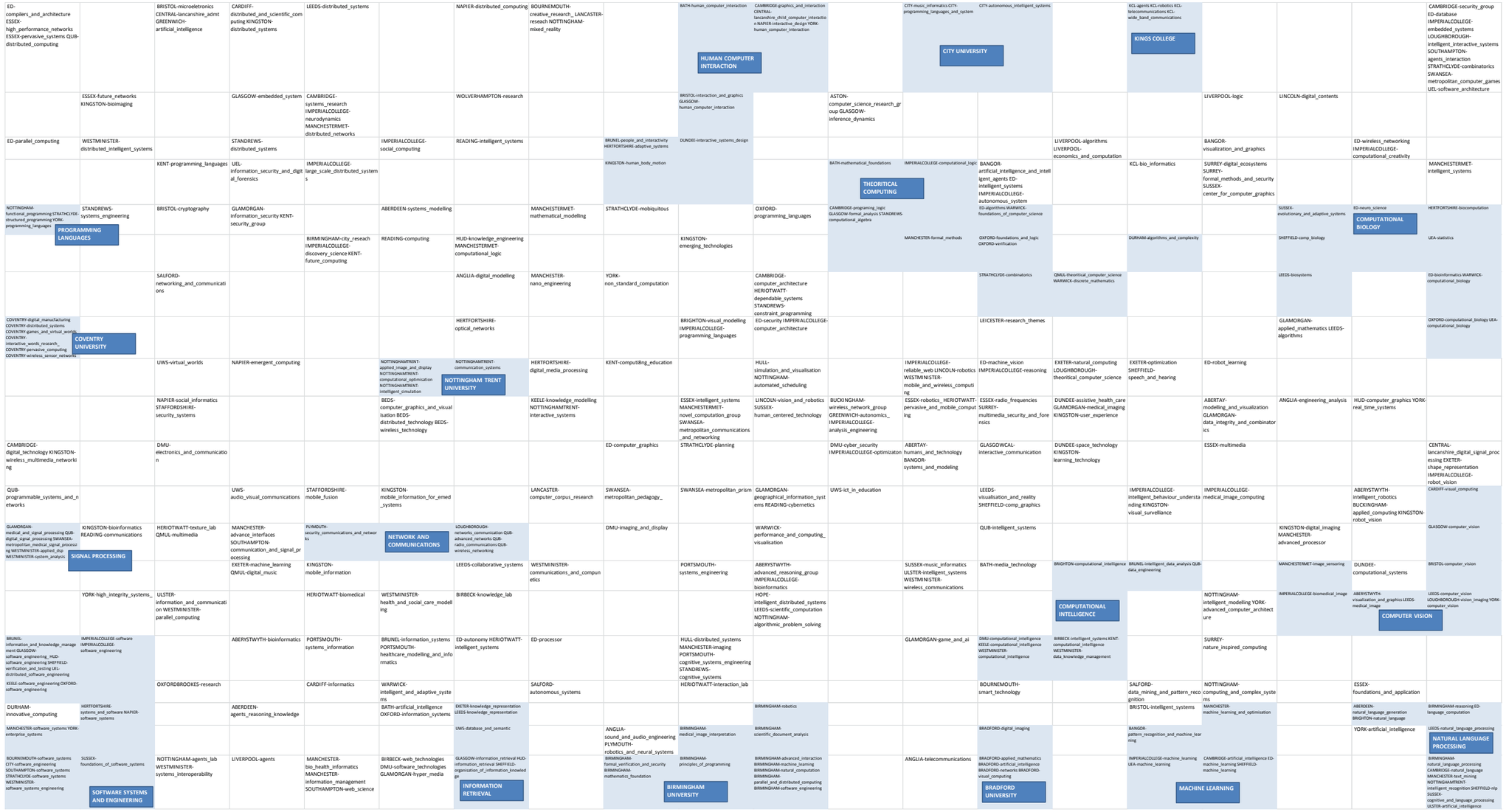

Figure 6.3 Topological Ordering of UK Computer Science research groups with SOMs and cluster themes based on the names of research groups

6.1.3. PCA grouping of research groups

The grouping algorithm simply assigns a research group to a particular cluster if its PCA component is above a defined threshold. This algorithm allows research groups to belong to more than one cluster, a feature that is important because research groups may have several themes, which should be taken into account if the similarity between two departments is determined by research groups' research areas.

Simple Clustering Algorithm

```
1 Clustering(threshold)
2 For each component as cluster i
3     If (ith Principal Component of RG > threshold)
4         Add the research group to cluster i
5 Delete clusters with less than 2 research groups
6 Assign any unassigned research group to its most similar
  cluster
```

Table 6.6 The influence of the value of threshold on the final groupings.

Threshold	No. of clusters	No. of research groups assigned to clusters in first instance	Inter-cluster similarity	Intra-cluster similarity
0.90	2	10	0.036	0.180
0.85	2	11	0.035	0.182
0.80	2	14	0.036	0.181
0.75	2	14	0.036	0.181
0.70	2	16	0.036	0.182
0.65	4	21	0.184	0.250
0.60	5	27	0.165	0.272
0.55	6	33	0.125	0.303
0.50	8	43	0.071	0.327
0.45	14	66	0.056	0.438
0.40	16	83	0.055	0.447
0.35	24	115	0.035	0.524
0.30	37	175	0.028	0.606
0.25	61	248	0.034	0.630
0.20	100	332	0.035	0.600

Table 6.6 shows clustering solutions from the simple clustering algorithm with varying threshold values. Threshold was in the range [min max] of the largest values of each PCA component.

A good clustering solution is when samples in the same cluster are very similar (intra cluster similarity) and the similarity between samples across clusters is low (inter cluster similarity). Inter and intra cluster similarities are used to evaluate the clustering solutions. Metrics to compute the similarity between cluster samples and clusters were

described in section 2.4.2.1.2 (page 56). In this study, intra cluster similarity is determined by the average similarity between samples in the same cluster and their corresponding cluster centres, while inter cluster similarity is determined by average similarity between all cluster centres.

$$Center\ Cluster_i = Average\ of\ all\ research\ group\ vectors\ in\ Cluster_i$$

$$Inter\ cluster\ similarity = Average\ cosine\ similarity\ between\ all\ cluster\ centers$$

$$Intra\ cluster\ similarity = Average\ cosine\ similarity\ of\ all\ research\ groups\ with\ their\ corresponding\ cluster\ center$$

The most appropriate threshold for the clustering task in this study was determined to be 0.30. From results in **Table 6.6**, it shows that threshold lower than 0.30 results in poorer clustering solutions in terms of both intra-cluster similarity and inter-cluster similarity.

A research group may belong to more than one cluster so the degree of membership to a particular cluster is computed by:

$$Research\ group\ membership\ in\ cluster(x) = \frac{C_x}{\sum_{i=1}^{n} C_i}$$

The numerator is the value of the principal component of the research group in cluster x, and the denominator is the sum of the principal components of the research group in all clusters the research group belongs to. Cluster membership is 1 when a research group appears in exactly one cluster.

The clustering results in **Figure 6.3** and **Figure 6.4** have used factor analysis to associate research groups to a particular research area. Previously, factor analysis has been used to associate authors with information science fields (Zhao and Strotmann, 2008). The visualizations in **Figure 6.3** and **Figure 6.4** show the clustering results based on factor analysis, with the names of the groups determined by their dominant members. In several cases, there was no obvious pattern that could assist in determining the theme of a group. Not all 37 groups in **Figure 6.4** seem to have a meaningful theme. Also, some groups can be merged or separated for a better clustering solution. However, the clustering result is the best that can be achieved with the algorithm used in this paper, in terms of inter-cluster and intra-cluster similarities. The identified groups can be seen as the fundamental research areas in UK computer science research, so these types of analysis may be useful for monitoring research disciplines.

Figure 6.4 shows the clustering solution with the PCA algorithm and the degree of membership of each research group in a particular cluster. The theme of each cluster are identical to the themes identified with SOMs, thus validates the accuracy of the algorithm. For example, from **Figure 6.3** and **Figure 6.4** research themes in software engineering, natural language processing, machine learning, computational intelligence, human computer interaction and computer

vision are some identified clusters on the SOM that were also identified by the PCA algorithm.

As keywords can give context to the research methods used in a particular research group, research groups in different specialities will be co-clustered if they use similar research concepts. Concepts used in multiple disciplines can be identified through these types of analyses, which in turn may foster inter-disciplinary collaboration. An example is the machine learning/natural language processing clusters in **Figure 6.4** that had groups from machine learning, natural language processing and statistics. These are different research areas that benefit from collaborating together. Larger scale analyses, not restricted to computer science may show different specialities that use similar research methods. This may help policy makers or heads of research groups decide if it would be beneficial to merge research groups to form research centres, because they will at least have some common ground.

6.1.4. Evaluation

It is difficult to evaluate the accuracy of a clustering solution because it consists of multiple groups and the best results from **Table 6.6** has up to 37 clusters which makes meaningful analysis difficult. To resolve this problem, the clustering solution was used to estimate the similarity between each pairs of departments and then the similarities were

compared with human similarity estimates. Even though human estimates on similarities is different from groupings, it can be used to evaluate the value of the clustering solution for a particular webometric task. Computing similarities between organisations can be useful some webometric studies. The similarity between two departments was estimated by the extent to which they co-occur in the same clusters, identified with the PCA clustering algorithm.

$$similarity\_between\ deparments(i,j) = \frac{No\ of\ RGs\ from\ i\ that\ appear\ in\ a\ cluster\ that\ contains\ a\ RG\ from\ j}{Number\ of\ research\ groups\ in\ i}$$

The similarity between two departments is a number between 0 and 1. This formula is asymmetric, that is, $similarity\_between(i,j) \neq similarity\_between(j,i)$. This is because departments may have different number of research groups. For example, suppose that all research groups in department A are co-clustered with research groups in department B, but department B still has additional research groups not co-clustered with department A. Then $similarity\_between(A,B) = 1$ whereas the $similarity\_between(B,A) < 1$.

The computed similarities between 10 random UK university computer science departments based on the similarity formula are shown in **Table 6.7**. A spreadsheet containing the computed similarities for all 76 universities can be found in http://goo.gl/MD3n1r.

Other similarity metrics like Euclidean distance or cosine similarity are alternatives that can be used to compute the similarities between research groups or departments, and dimension reduction through factor analysis can improve computation time. Thresholding PCA components has also been used in bibliometric analysis (Schreiber, Malesios and Psarakis, 2012; Perianes-Rodríguez, Olmeda-Gómez and Moya-Anegón, 2009).

Table 6.7 Similarity between 10 random UK computer science departments based on the co-occurrence of their research groups in clusters identified by clustering with PCA.

	hope	**bath**	**herts**	**qmul**	**liv**	**nott**	**gcal**	**man**	**brunel**	**mmu**
hope	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
bath	0.00	1.00	0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00
herts	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.20	0.40	0.00
qmul	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.67	0.00	0.00
liv	0.00	0.00	0.00	0.00	1.00	0.25	0.00	0.25	0.00	0.75
nott	0.08	0.29	0.00	0.00	0.14	1.00	0.00	0.43	0.07	0.14
gcal	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
man	0.00	0.20	0.10	0.20	0.10	0.30	0.00	1.00	0.10	0.10
brunel	0.00	0.00	0.38	0.00	0.00	0.37	0.00	0.25	1.00	0.25
mmu	0.00	0.00	0.00	0.00	0.17	0.17	0.00	0.17	0.17	1.00

In order to determine the extent to which the computed results in **Table 6.7** genuinely reflect inter-department research similarity, two independent coders with computing backgrounds compared the websites of the research groups and assigned a number on a scale of 0 – 10, for the similarity between the departments' research, as described on their websites. The coders were given the following instructions:

- Based on the websites of the research groups in two departments, on a scale of 0 – 10, indicate the extent to which the research in department 1 is part of the research in department 2? Zero being when there is no common research interest and ten being when all research areas in department 1 is part or all of the research areas in department 2.

The comparisons between research groups is based on the opinions of the coders. It can be time consuming to compare the web pages from research groups of all 76 universities, which is why 10 universities were randomly selected for the comparisons. Krippendorff's alpha can be used to determine the reliability of the coding results.

Krippendorff's alpha (α) (Krippendorff, 2004) was used to compute inter-coder reliability. It is a method that is widely used (Stuart, Thelwall and Harries, 2007; Thelwall, Buckley, et al., 2010; Ozel and Park, 2011; Tuomaala, Järvelin and Vakkari, 2014). Using the

interval metric to determine the difference in the value of variables, α was 0.9020, which suggests high agreement between the coders. In order to get a reliable similarity estimate, the average of the two coders was used as the gold standard.

The mean absolute error between the two coders was 0.94, while the mean absolute error between the gold standard and the values in **Table 6.7** (multiplied by 10 to give them the same scale as the gold standard) was 1.254. The low difference in mean absolute error between the human coders and the gold standard vs. computed similarity suggests that the clustering results are not substantially less accurate than human judgements, based upon the contents of the departmental websites.

The computer science key phrases in the research group homepages from the university departments in **Table 6.7** were merged into 10 documents, depending on which university the research group belonged to. The cosine similarities between the tf-idf vectors of the 10 universities were then computed. Cosine similarity is one of the most popular metrics used to measure the similarity between text documents (Huang, 2008). The absolute error between the gold standard and cosine similarity was 1.007, which is marginally better than the method used in this research and so the proposed new method does not outperform existing standard metrics. However, the new method used in this research essentially estimates the percentage of a department's

research that is also investigated in another university, an asymmetric approach. In contrast, cosine similarity and other traditional similarity metrics are symmetric. Traditional metrics may therefore not show that smaller departments are 'similar' to much larger departments in the sense that much research carried out in the smaller department is also researched in the larger department. Hence the new method may be useful when this overlap is more important than overall similarity, which seems likely to be true for the task of finding research partners.

Group 1		Group 5		Group 9		Group 13		Group 20		Group 29	
CAMBRIDGE-security group	1.00	SHEFFIELD-machine learning	1.00	DURHAM-innovative computing	1.00	OXFORD-computational biology	1.00	BEDS-wireless technology	1.00	MANCHESTER-information management	1.00
ED-database	1.00	ED-machine learning	1.00	KEELE-software engineering	1.00	ED-robot learning	1.00	NOTTINGHAMTRENT-intelligent simulation	1.00	SALFORD-networking and communications	1.00
IMPERIALCOLLEGE-embedded systems	1.00	BIRMINGHAM-natural language processing	1.00	CITY-software engineering	1.00	EXETER-machine learning	1.00	HERTFORTSHIRE-optical networks	1.00	BIRBECK-knowledge lab	1.00
LOUGHBOROUGH-intelligent interactive systems	1.00	SHEFFIELD-nlp	1.00	MANCHESTER-software systems	1.00	GLAMORGAN-applied mathematics	1.00	KEELE-knowledge modelling	1.00	BIRBECK-web technologies	1.00
SOUTHAMPTON-agents interaction	1.00	CAMBRIDGE-natural language	1.00	HUD-software engineering	1.00	IMPERIALCOLLEGE-biomedical image	1.00	LANCASTER-computer corpus research	1.00	GLAMORGAN-information security	1.00
SWANSEA-metropolitan computer games	1.00	MANCHESTER-text mining	1.00	STRATHCLYDE-software systems	1.00	IMPERIALCOLLEGE-machine learning	1.00	NOTTINGHAMTRENT-communication systems	1.00	LEEDS-collaborative systems	1.00
UEL-software architecture	1.00	ULSTER-artificial intelligence	1.00	OXFORD-software engineering	1.00	MANCHESTER-machine learning and optimisation	1.00	NOTTINGHAMTRENT-computational optimisation	1.00	LEICESTER-research themes	1.00
ED-wireless networking	1.00	SUSSEX-cognitive and language processing	1.00	WESTMINISTER-software systems engineering	1.00	MANCHESTERMET-intelligent systems	1.00	SWANSEA-metropolitan pedagogy	1.00	LOUGHBOROUGH-networks communication	1.00
IMPERIALCOLLEGE-computational creativity	1.00	LEEDS-natural language processing	1.00	BOURNEMOUTH-software systems	1.00	NOTTINGHAM-computing and complex systems	1.00	UWS-ict in education	1.00	QUB-advanced networks	1.00
LINCOLN-digital contents	1.00	NOTTINGHAMTRENT-intelligent recognition	1.00	NAPIER-software systems	1.00	SALFORD-data mining and pattern recognition	1.00	WESTMINISTER-wireless communications	1.00	QUB-radio communications	1.00
LEEDS-distributed systems	1.00	CAMBRIDGE-artificial intelligence	1.00	YORK-enterprise systems	1.00	UEA-computational biology	1.00	NOTTINGHAMTRENT-interactive systems	0.54	QUB-wireless networking	1.00
WOLVERHAMPTON-research	1.00	YORK-artificial intelligence	1.00	SOUTHAMPTON-software systems	1.00	YORK-advanced computer architecture	1.00	BEDS-distributed technology	0.54	STANDREWS-distributed systems	1.00
ESSEX-robotics	1.00	BIRMINGHAM-reasoning	1.00	BRUNEL-information and knowledge management	1.00	BRISTOL-intelligent systems	0.52	BEDS-computer graphics and visualisation	0.53	SUSSEX-center for computer graphics	1.00
ABERTAY-humans and technology	1.00	ED-language computation	1.00	UEL-distributed software engineering	1.00	WARWICK-computational biology	0.49	Group 21		UEL-information security and digital forensics	1.00
BRISTOL-cryptography	1.00	ED-bioinformatics	1.00	GLASGOW-software engineering	1.00	YORK-computer vision	0.49	PORTSMOUTH-cognitive systems engineering	1.00	WARWICK-intelligent and adaptive systems	1.00
CAMBRIDGE-systems research	1.00	ABERDEEN-natural language generation	1.00	SHEFFIELD-verification and testing	1.00	BANGOR-pattern recognition and machine learning	0.49	MANCHESTER-imaging	1.00	Group 30	
ESSEX-intelligent systems	1.00	BRIGHTON-natural language	1.00	SUSSEX-foundations of software systems	1.00	UEA-statistics	0.45	HULL-distributed systems	1.00	QUB-intelligent systems	1.00
ESSEX-multimedia	1.00	ESSEX-foundations and application	1.00	IMPERIALCOLLEGE-software engineering	1.00	UEA-machine learning	0.45	STANDREWS-cognitive systems	1.00	BOURNEMOUTH-creative research	1.00
GREENWICH-autonomics	1.00	SHEFFIELD-speech and hearing	1.00	IMPERIALCOLLEGE-software	1.00	GLASGOW-inference dynamics	0.45	SURREY-digital ecosystems	1.00	BRADFORD-digital imaging	1.00
HULL-simulation and visualisation	1.00	UEA-machine learning	0.55	HERTFORTSHIRE-systems and software	1.00	SURREY-nature inspired computing	0.39	HERIOTWATT-dependable systems	1.00	HERTFORTSHIRE-digital media processing	1.00
IMPERIALCOLLEGE-neurodynamics	1.00	UEA-statistics	0.55	OXFORDBROOKES-research	1.00	Group 14		HERIOTWATT-interaction lab	1.00	SHEFFIELD-comp graphics	0.57
IMPERIALCOLLEGE-programming languages	1.00	BANGOR-pattern recognition and machine learning	0.51	ABERYSTWYTH-bioinformatics	1.00	BIRMINGHAM-natural computation	1.00	PORTSMOUTH-systems engineering	1.00	LEEDS-visualisation and reality	0.56
KENT-computing education	1.00	WARWICK-computational biology	0.51	NAPIER-emergent computing	1.00	BIRMINGHAM-advanced interaction	1.00	SURREY-formal methods and security	1.00	NOTTINGHAMTRENT-interactive systems	0.54
KENT-programming languages	1.00	BRISTOL-intelligent systems	0.48	ULSTER-information and communication	1.00	BIRMINGHAM-robotics	1.00	SURREY-multimedia security and forensics	1.00	BRUNEL-people and interactivity	0.50
LIVERPOOL-algorithms	1.00	BIRMINGHAM-machine learning	0.44	WESTMINISTER-parallel computing	1.00	BIRMINGHAM-scientific document analysis	1.00	WESTMINISTER-communications and compunetics	1.00	Group 31	
LIVERPOOL-economics and computation	1.00	Group 6		WESTMINISTER-systems interoperability	1.00	BIRMINGHAM-parallel and distributed computing	1.00	SURREY-nature inspired computing	0.39	DUNDEE-assistive health care	1.00
LIVERPOOL-logic	1.00	QUB-digital signal processing	1.00	YORK-high integrity systems	1.00	BIRMINGHAM-principles of programming	1.00	Group 22		DUNDEE-interactive systems design	1.00
MANCHESTERMET-distributed networks	1.00	GLAMORGAN-medical and signal processing	1.00	BIRMINGHAM-software engineering	0.40	BIRMINGHAM-medical image interpretation	1.00	HERIOTWATT-pervasive and mobile computing	1.00	DUNDEE-space technology	1.00
STRATHCLYDE-planning	1.00	WESTMINISTER-applied dsp	1.00	Group 10		BIRMINGHAM-mathematics foundation	1.00	ESSEX-radio frequencies	1.00	MANCHESTER-formal methods	1.00
SUSSEX-human centered technology	1.00	WESTMINISTER-system analysis	1.00	GLASGOW-computer vision	1.00	ANGLIA-engineering analysis	1.00	READING-intelligent systems	1.00	STRATHCLYDE-mobiquitous	1.00
SWANSEA-metropolitan communications and networking	1.00	SWANSEA-metropolitan medical signal processing	1.00	LEEDS-computer vision	1.00	ANGLIA-sound and audio engineering	1.00	SHEFFIELD-organisation of information knowledge	1.00	BATH-mathematical foundations	1.00
SWANSEA-metropolitan prism	1.00	QUB-programmable systems and networks	1.00	BRISTOL-computer vision	1.00	ANGLIA-telecommunications	1.00	GLASGOW-information retrieval	0.45	ED-machine vision	1.00
STRATHCLYDE-combinatorics	0.99	READING-communications	1.00	CENTRAL-lancanshire digital signal processing	1.00	BIRMINGHAM-formal verification and security	1.00	HUD-information retrieval	0.44	GLASGOW-formal analysis	1.00
ED-compilers and architecture	0.34	KINGSTON-bioinformatics	1.00	LOUGHBOROUGH-vision imaging	1.00	OXFORD-foundations and logic	1.00	Group 23		IMPERIALCOLLEGE-reliable web	1.00
GREENWICH-artificial intelligence	0.20	KCL-telecommunications	1.00	ABERYSTWYTH-visualization and graphics	1.00	PLYMOUTH-robotics and neural systems	1.00	CAMBRIDGE-computer architecture	0.43	KINGSTON-learning technology	1.00
Group 2		QMUL-multimedia	1.00	KINGSTON-robot vision	1.00	BIRMINGHAM-software engineering	0.60	IMPERIALCOLLEGE-computer architecture	0.42	OXFORD-verification	1.00
COVENTRY-pervasive computing	1.00	KCL-robotics	1.00	ABERTAY-modelling and visualization	1.00	BIRMINGHAM-machine learning	0.56	Group 24		WESTMINISTER-mobile and wireless computing	1.00
COVENTRY-wireless sensor networks	1.00	KINGSTON-wireless multimedia networking	1.00	ABERYSTWYTH-intelligent robotics	1.00	Group 15		LINCOLN-vision and robotics	0.49	Group 32	
COVENTRY-distributed systems	1.00	GLAMORGAN-game and ai	1.00	BATH-media technology	1.00	ED-parallel computing	1.00	LINCOLN-robotics	0.49	READING-cybernetics	0.50
COVENTRY-games and virtual worlds	1.00	CAMBRIDGE-digital technology	1.00	BUCKINGHAM-applied computing	1.00	ESSEX-future networks	1.00	Group 25		GLAMORGAN-medical imaging	0.50
COVENTRY-digital manufacturing	1.00	STAFFORDSHIRE-mobile fusion	1.00	CARDIFF-visual computing	1.00	GLASGOW-embedded system	1.00	LINCOLN-vision and robotics	0.51	Group 33	
COVENTRY-interactive words research	1.00	HERIOTWATT-texture lab	1.00	DMU-imaging and display	1.00	WESTMINISTER-distributed intelligent systems	1.00	LINCOLN-robotics	0.51	READING-cybernetics	0.50
BRIGHTON-computational intelligence	1.00	KCL-bio informatics	1.00	DUNDEE-computational systems	1.00	KINGSTON-bioimaging	0.54	Group 26		GLAMORGAN-medical imaging	0.50
GLASGOWCAL-interactive communication	1.00	KCL-wide band communications	1.00	DURHAM-algorithms and complexity	1.00	ESSEX-high performance networks	0.48	BIRMINGHAM-city research	1.00	Group 34	
NAPIER-distributed computing	1.00	KINGSTON-mobile information	1.00	ED-algorithms	1.00	QUB-distributed computing	0.44	KENT-future computing	1.00	MANCHESTERMET-computational logic	1.00
NAPIER-social informatics	1.00	MANCHESTER-advance interfaces	1.00	HUD-computer graphics	1.00	CENTRAL-lancanshire admt	0.43	IMPERIALCOLLEGE-discovery science	1.00	HUD-knowledge engineering	1.00
UWS-virtual worlds	1.00	PLYMOUTH-security communications and networks	1.00	IMPERIALCOLLEGE-medical image computing	1.00	ED-compilers and architecture	0.31	STANDREWS-systems engineering	1.00	MANCHESTERMET-mathematical modelling	1.00
Group 3		SOUTHAMPTON-communication and signal processing	1.00	IMPERIALCOLLEGE-robot vision	1.00	BRISTOL-microeletronics	0.22	ANGLIA-digital modelling	1.00	ABERDEEN-systems modelling	1.00
GLAMORGAN-hyper media	1.00	UWS-audio visual communications	1.00	LEEDS-medical image	1.00	GREENWICH-artificial intelligence	0.20	IMPERIALCOLLEGE-large scale distributed systems	1.00	Group 35	
SOUTHAMPTON-web science	1.00	ESSEX-pervasive systems	0.59	MANCHESTER-advanced processor	1.00	Group 16		KENT-security group	1.00	ED-computer graphics	1.00
MANCHESTER-bio health informatics	1.00	BRISTOL-microeletronics	0.22	NOTTINGHAM-intelligent modelling	1.00	STRATHCLYDE-structured programming	1.00	YORK-programming languages	0.50	KINGSTON-mobile information for emed systems	1.00
LIVERPOOL-agents	1.00	Group 7		YORK-real time systems	1.00	BRADFORD-artificial intelligence	1.00	Group 27		MANCHESTERMET-image sensoring	1.00
PORTSMOUTH-healthcare modelling and informatics	1.00	WESTMINISTER-computational intelligence	1.00	YORK-computer vision	0.51	NOTTINGHAM-functional programming	1.00	LEEDS-scientific computation	1.00	BEDS-computer graphics and visualisation	0.47
BANGOR-artificial intelligence and intelligent agents	1.00	BIRBECK-intelligent systems	1.00	CENTRAL-lancanshire admt	0.43	BRIGHTON-visual modelling	1.00	IMPERIALCOLLEGE-bioinformatics	1.00	BEDS-distributed technology	0.46
CARDIFF-informatics	1.00	READING-computing	1.00	Group 11		BANGOR-systems and modeling	1.00	BRUNEL-information systems	1.00	Group 36	
CITY-autonomous intelligent systems	1.00	BOURNEMOUTH-smart technology	1.00	BATH-human computer interaction	1.00	BRADFORD-networks	1.00	GLAMORGAN-geographical information systems	1.00	LEEDS-knowledge representation	1.00
DMU-software technologies	1.00	DMU-computational intelligence	1.00	CENTRAL-lancanshire child computer interaction	1.00	BRADFORD-visual computing	1.00	LOUGHBOROUGH-theoritical computer science	1.00	EXETER-knowledge representation	1.00
IMPERIALCOLLEGE-autonomous system	1.00	KEELE-computational intelligence	1.00	CAMBRIDGE-graphics and interaction	1.00	CAMBRIDGE-programing logic	1.00	WESTMINISTER-health and social care modelling	1.00	BATH-artificial intelligence	1.00
IMPERIALCOLLEGE-computational logic	1.00	KENT-computational intelligence	1.00	NOTTINGHAMTRENT-applied image and display	1.00	IMPERIALCOLLEGE-social computing	1.00	BRUNEL-people and interactivity	0.50	IMPERIALCOLLEGE-intelligent behaviour understanding	1.00
KCL-agents	1.00	QUB-data engineering	1.00	ASTON-computer science research group	1.00	OXFORD-programming languages	1.00	NOTTINGHAM-algorithmic problem solving	0.47	OXFORD-information systems	1.00
NOTTINGHAM-agents lab	1.00	ULSTER-intelligent systems	1.00	BRISTOL-interaction and graphics	1.00	STAFFORDSHIRE-security systems	1.00	LEEDS-visualisation and reality	0.44	SALFORD-autonomous systems	1.00
PORTSMOUTH-systems information	1.00	WESTMINISTER-data knowledge management	1.00	GLASGOW-human computer interaction	1.00	YORK-programming languages	0.50	SHEFFIELD-comp graphics	0.43	STANDREWS-constraint programming	1.00
STANDREWS-computational algebra	1.00	QUB-distributed computing	0.56	KINGSTON-user experience	1.00	CITY-programming languages and system	0.30	Group 28		SUSSEX-music informatics	1.00
UWS-database and semantic	1.00	KINGSTON-bioimaging	0.54	LANCASTER-reseach	1.00	Group 17		SUSSEX-evolutionary and adaptive systems	1.00	HUD-information retrieval	0.56
YORK-non standard computation	1.00	ESSEX-high performance networks	0.52	NOTTINGHAM-mixed reality	1.00	HOPE-intelligent distributed systems	1.00	ED-neuro science	1.00	GLASGOW-information retrieval	0.55
CITY-music informatics	0.51	ESSEX-pervasive systems	0.41	YORK-human computer interaction	0.52	ABERYSTWYTH-advanced reasoning group	1.00	BANGOR-visualization and graphics	1.00	CITY-music informatics	0.49
NAPIER-interactive design	0.48	ED-compilers and architecture	0.35	NAPIER-interactive design	0.52	NOTTINGHAM-algorithmic problem solving	0.53	HERTFORTSHIRE-adaptive systems	1.00	ABERDEEN-agents reasoning knowledge	0.06
YORK-human computer interaction	0.48	BRISTOL-microeletronics	0.22	GLASGOW-inference dynamics	0.45	Group 18		HERTFORTSHIRE-biocomputation	1.00	Group 37	
CITY-programming languages and system	0.30	GREENWICH-artificial intelligence	0.20	CITY-programming languages and system	0.30	KINGSTON-human body motion	1.00	SHEFFIELD-comp biology	1.00	WARWICK-discrete mathematics	1.00
ABERDEEN-agents reasoning knowledge	0.06	Group 8		Group 12		KINGSTON-digital imaging	0.54			STRATHCLYDE-combinatorics	0.01
Group 4		MANCHESTERMET-novel computation group	1.00	MANCHESTER-nano engineering	1.00	KINGSTON-visual surveillance	0.50			BRADFORD-applied mathematics	1.00
ED-processor	1.00	IMPERIALCOLLEGE-analysis engineering	1.00	QMUL-digital music	1.00	Group 19				ED-intelligent systems	1.00
ED-security	1.00	EXETER-shape representation	1.00	BUCKINGHAM-wireless network group	1.00	NOTTINGHAM-automated scheduling	1.00			GLAMORGAN-data integrity and combinatorics	1.00
IMPERIALCOLLEGE-reasoning	1.00	WARWICK-performance and computing visualisation	1.00	CARDIFF-distributed and scientific computing	1.00	KINGSTON-visual surveillance	0.53			LEEDS-algorithms	1.00
QMUL-theoritical computer science	1.00	BRUNEL-intelligent data analysis	1.00	ED-autonomy	1.00	KINGSTON-digital imaging	0.46			WARWICK-foundations of computer science	1.00
IMPERIALCOLLEGE-computer architecture	0.58	DMU-cyber security	1.00	HERIOTWATT-intelligent systems	1.00						
CAMBRIDGE-computer architecture	0.57	DMU-electronics and communication	1.00	IMPERIALCOLLEGE-optimizaton	1.00						
		EXETER-natural computing	1.00	KINGSTON-emerging technologies	1.00						
		EXETER-optimization	1.00								
		HERIOTWATT-biomedical	1.00								
		KINGSTON-distributed systems	1.00								

Figure 6.4 Result of grouping UK computer science research groups with PCA clustering algorithm and the degree of membership to each cluster.

6.2. Conclusion

This chapter used unsupervised machine learning techniques to cluster computer science research groups in UK universities based upon the text in their homepages and showed that machine learning techniques can generate a clustering solution that is reasonable in the sense that inter-departmental similarities inferred from it are not greatly different from human judgements based on the departmental websites.

This study assumes that all relevant research information about a computer science department can be identified from its website. However, websites and URLs change frequently and so some homepages used in this study may no longer be available, and the content of the homepages may also have changed. For example, Oxford Brookes computer science research groups did not have dedicated homepages when the data was collected, but do now. The study is also restricted to one type of department and one country and so the results may well be less good for other disciplines and countries that use the web significantly less.

The components from the PCA of computer science term frequencies in UK universities' research group homepages were clustered with self-organising maps. The results showed that the PCA is not only useful for dimension reduction for this type of webometric data

but significantly improves the speed of the SOM, because of the reduced dimension of vectors. It also improves the technical accuracy of the clustering solution in terms of the quantization and topographical error. The quantization and topographical error of SOM is significantly less than the error with tf-idf vectors.

From the visualization of the SOM, specific PCA components seemed to be directly responsible for particular clusters identified on the SOM, and so a simple clustering method based on PCA components was used to group research groups. The similarity between the computer science departments that the research group belonged to, was then computed based on how often they co-occurred in the same clusters.

The grouping of research groups based on the PCA components were identical to those in the SOM, which supports the use of PCA for feature extraction in machine learning contexts. The resulting similarity values from the clustering technique seem to be a reliable indicator of the similarity between computer science departments' research interest, as the mean absolute error between two human coders was only marginally lower than the mean absolute error of the computed similarities and the gold standard.

From the size of the clusters, the distribution of key computer science research areas may be identified, which may be useful for observing the computer science research field. Monitoring research

fields is an application area for webometrics research methods (Thelwall, Klitkou, et al., 2010) and machine learning methods may now be used for this task with some confidence.

Although Thijs and Glänzel (2010) found that among collaborating institutions there was only a weak association between the similarity of their research profiles and their collaboration intensity, the study also showed that an institution's research profile was more similar to those of its collaborating partners than to those of other institutions. Hence, if the similarity between universities is a factor that can influence the choice of a collaborator, a full scale study using the methods described in this paper may also help a little to identify universities that may benefit from collaboration. This could assist policy makers looking for promising future collaborators with shared research interests, or seeking to identify groups that can be merged to foster interdisciplinary collaboration.

7. Conclusions

Until now, few webometric studies have used machine learning techniques. This thesis fills this gap with a series of webometrics studies that use supervised and unsupervised learning techniques (a) in order to better utilise web based data to estimate the extent to which a university collaborates and (b) for exploratory cluster analysis to aid policy makers to track changes in research fields, and to help policy makers to choose future suitable collaborators.

7.1. Limitations

In order to automatically classify web pages, a random selection of web pages were selected and manually classified to form the training and test set. Even though classification was done systematically, using the main missions of a university as a guide and final categories verified by another researcher, web pages were classified by a single researcher, thus making the resulting test and training set largely subjective. The automatic classification model always misclassified certain web page types because they contribute only a small percentage to the total number of web pages in a universities' website. This puts a question mark on the suitability of the web page types in **Table 4.1** and perhaps a more adequate classification scheme can be developed.

The parameters of the machine learning algorithms were not systematically tweaked to identify the optimal settings for each algorithm, so a study that takes the settings of the machine learning parameters into account may achieve higher accuracy than the results reported in **Table 4.2** and **Table 4.9**.

Because of the difficulty in obtaining data from the Web of Science or other ISI databases, co-authorship data was limited to only the 36 UK universities that appeared in the CWTS 2013 Leiden ranking and the research project data in chapter 5 did not cover all the research projects awarded to UK universities. Although this is a limitation, it emphasises the importance of part of the goals of this thesis in attempting to make hyperlink data more reliable as an indicator for collaboration with machine learning methods, which can in turn serve as an alternate data source for those researchers without access to co-authorship data from publication databases.

Correlation tests were used to investigate the association between variables but results should be interpreted with caution because correlation does not imply causation, which is why it has been advised that data should also be manually checked in order to give contexts to webometrics results. However using machine learning techniques to previously pre-process data before the correlation tests increases the validity of correlation results.

7.2. Key findings

7.2.1. Findings from automatically classifying web pages and hyperlink targets

The first study in Chapter 4 described machine learning methods that classified web pages into one of 8 categories (see **Table 4.1** page 93) that university web pages could belong to. Up to 78% accuracy was achieved with support vector machines (see **Table 4.2** page 98) using the top 500 features with the highest information gain because the machine learning algorithms were less accurate accuracy when more features were used (see **Figure 4.2** page 100).

The target web page category can also be determined from the information in the source page and some outlinks in different web page types have unique characteristics. Support vector machines also produced the best result; 74% (see **Table 4.9** page 109) in identifying target web page types. This means that the techniques used in this study (Chapter 4) can be used to address the problem stated in previous webometrics studies that methods for automatic classification of links in university web sites is needed to fully harness the potential of

hyperlinks for collaboration studies (Stuart, Thelwall and Harries, 2007), because both the source and target web page type that are needed for link classification can be automatically identified with relatively high accuracy. Human inter coder agreement in the classification of hyperlinks is between 70% and 98% depending on the complexity of the classification scheme (Holmberg, 2009). But in quantitative webometrics research, the gain of automation is greater than the cost of few misclassifications. Automatic classification of university web page types seems therefore to be accurate enough to aid in more efficient hyperlink classification on a large scale, if the reasons for link creation can be inferred from the connection between web page types.

In terms of the accuracy of each machine learning technique, machine learning techniques that use complex mathematical functions produce more accurate results at the cost of the clarity of the resulting classifiers. SVMs produced better results than the other supervised learning algorithms in Section 2.4.1. Decision trees are less accurate than other models (approximately 6% less accurate than support vector machines); but the classification model is a set of rules that can be easily explained. Automatic classification can also be used to design more efficient focused crawlers for webometrics studies.

7.2.2. Findings from using machine learning to filter out irrelevant links for collaboration studies

Although previous webometrics research has attempted to use hyperlinks as indicators for collaboration, it had not previously exploited with manual or automatic link classification.

The majority of hyperlinks in academic websites are not created because of collaboration, so if those links that are not created because of collaboration reasons can be automatically identified and excluded; it increases the suitability for using hyperlink data as a collaboration indicator.

The study (Section 5.1) automatically identified academic links in two facets: links between university staff webpages and all links between university web pages excluding those from the support web page category (web pages created to provide services) using support vector machines; with up to 95% accuracy, and then investigated if these links associate better with collaboration, because the results from the study in Chapter 4 showed that the links between university staff-related web pages are more likely to be created because of collaboration reasons than links from other web page types (see **Table**

4.10 page 112) and support pages very likely do not contain links created for collaboration reasons (Stuart, Thelwall and Harries, 2007).

Academic links showed increased association with the extent a university collaborates with all other universities (spearman correlation 0.75) than when all links were used (spearman correlation 0.57), but there was only marginal improvement in the extent two universities collaborate together when using only academic links (spearman correlation 0.316) rather than all links (spearman correlation 0.271) (see **Table 5.1** page 129 and **Table 5.3** page 140). This may have been because the methods used to identify academic links do not adequately filter the majority of links created for non-collaborative reasons, and also, some collaboration between universities that may be identified through analysis of hyperlink data may not be identified in tradition data used in collaboration studies.

Results from this study highlight the suitability of using machine learning methods for more effective link analysis in large scale studies. Also the results, particularly the limited correlation between inter academic links and inter university collaboration emphasises the need for improvement in the methods used in this study for more accurate results.

7.2.3. Findings from investigating web mentions as a collaboration indicator

Previous research has shown that there is significant correlation between links and organisation mentions and that organisation mentions can be used for impact studies (Thelwall and Sud, 2011). This study (Section 5.1.5) analysed through correlation the extent to which organisation name mentions can be used as an indicator for investigating inter-university collaboration compared to hyperlinks.

Links had higher correlation with collaboration (0.271) than name mentions had with collaboration (0.121) (see **Table 5.4** page 140) which suggests that links are slightly better as an indicator for collaboration than name mentions are. However, from the significant change in the normalized and un-normalized correlation results (see **Table 5.4** page 140 and **Table 5.5** page 142); it shows that the size of a university also has a significant influence on the number of mentions just as it has on the number of links. The reasons for name mentions in the 8 web page categories in **Table 4.1** (page 93) are also broadly the same as the reasons for hyperlink creation (see **Table 5.6** page 142), which suggests that machine learning techniques can also be used to increase the suitability for using name mentions as a collaboration

indicator just as it improved the quality hyperlink data in the study in section 5.1 (page 117). Perhaps name mentions and hyperlinks can be systematically combined in future webometrics research.

7.2.4. Findings from the Exploratory cluster analysis of computer science research groups

Text in organisations' websites have been shown to be able to identify related organisations through statistical techniques like multi-dimensional scaling (Vaughan and You, 2010). This study (see section 6.1 page 146) used the text in the homepages of UK computer science research groups for exploratory cluster analysis to identify if useful information can be extracted from the clustering solution.

Principal component analysis was used to reduce the dimension of input samples. The results of clustering showed that using PCA components instead of tf-idf vectors as input did not only speed up the clustering algorithm because of the reduced input dimensions, but also improved the clustering solution in terms of both quantization and topographical error; PCA [QE: 0.565 TE: 0.005] TF-IDF [QE: 0.957 TE: 0.040] (see **Table 6.1** to **Table 6.5** page 156).

From the topographical ordering of the research groups on the SOM (see **Figure 6.3** page 161), it showed that particular PCA components seemed directly responsible for specific clusters, so a clustering algorithm was used to simply assigned a research group to a cluster if its PCA component was above a certain threshold (see section 6.1.3 page 162). The algorithm allows for research groups to belong to more than one cluster, which is useful to determine the similarities of computer science departments based on cluster co-occurrence, because research groups can have multiple research areas.

The cosine inter-cluster similarity of PCA clustering algorithm was 0.028 and the cosine intra-cluster similarity was 0.606 (see **Table 6.6** page 163). Similarity could be in the range [0 1], and good clustering solutions should have high intra-cluster similarity and low inter-cluster similarity. The clusters identified with the PCA clustering algorithm were identical to the ones on the SOM, both having themes based on the different computer science research areas and in some cases clusters based on departments from the same institution (see **Figure 6.3** page 161 and **Figure 6.4** page 173). This means that the PCA clustering method used in this study may be useful for clustering purposes.

The similarity between two computer science departments was computed based on how often their research groups appeared in the same clusters (see **Table 6.7** page 169), and the absolute error

computed similarity and the gold standard is only marginally worse than the absolute error between human coders.

This technique may be beneficial to give information to young researchers when they decide future research interests, as the size of clusters may be used as an indicator for the key research areas. The clustering method can be further developed to a tool that policy makers can use to monitor emergent or dying research fields or identify possible future collaborators, if the similarity between institutions influences collaboration (Thijs and Glänzel, 2010). These have been listed as application areas of webometrics research (Thelwall, Klitkou, et al., 2010).

7.3. Contribution to knowledge

This thesis has used four studies to demonstrate that machine learning techniques can help webometrics research, particularly in large scale link analyses.

The need for automatic classification of hyperlinks was emphasized in (Stuart, Thelwall and Harries, 2007), so that hyperlink data can reach its full potential in webometrics research. Chapter 4 shows that university webpages can be classified with a reasonable degree of accuracy using machine learning methods, thus creating the

potential for more effective and efficient hyperlink analysis, or classification based on inter-page relationships in large scale webometrics studies, which in turn brings hyperlink data a step closer to reaching its full potential in webometrics research.

Research has shown that hyperlink data do not reflect collaboration ties among organisations (Shari, Haddow and Genoni, 2012; Kretschmer, Kretschmer and Kretschmer, 2007), however Kretschmer and her colleagues (2007) suggest that if the reasons for link creation is taken into account, hyperlink data may be better used for collaboration studies. Stuart, Thelwall and Harries (2007) also showed that links from certain web pages likely do not contain links that may have been created because of collaboration, but until now, few webometrics research have taken the web page type or hyperlink category into account when using hyperlink data for collaboration analyses. Results from the study in Section 5.1 show that using machine learning techniques to filter out some links that may not have been created for collaborative reasons can help to improve the quality of the hyperlink data for collaboration studies. The study describes methods that can be used to make hyperlink data a more suitable indicator for collaboration studies.

It has been shown over the years that the data used in webometrics research is not restricted to hyperlink data alone. Text and in particular organisation name mentions have been used in a number

of studies. Thelwall and Sud (2011) have shown that organisation name mentions can be used for impact studies and name mentions can be used as substitutes for inlinks in Spanish academic websites (Ortega, Orduña-Malea and Aguillo, 2013). However, few studies have investigated the suitability of name mentions for collaboration studies. Results from the study in section 5.1.5 shows that although link data are slightly better than name mentions as an indicator for collaboration, name mentions may also be used for these types of analyses. More importantly, name mentions can also be made more suitable for collaboration studies using the techniques described in Chapter 4 and in Section 5.1, because the reasons for name mentions in academic websites seem to be broadly the same as the reasons for link creation in academic websites.

The majority of early webometrics studies have analysed hyperlink relationships with graph based techniques. The final study in Section 6.1 is exploratory, in that it investigated what could be identified through clustering of the text in academic web pages. The study showed that the methods described could be used to develop a tool that may assist policy makers or department heads in monitoring research fields, identifying suitable future collaborators or merging research groups to from research centres which in turn may foster interdisciplinary collaboration. Some of these have been listed as

application areas of webometrics research (Thelwall, Klitkou, et al., 2010).

7.4. Further work

The classification scheme and data set were created and classified by a single researcher which is a limitation. Future studies should aim to develop a classification schemes for different types of webometrics studies that is agreed upon by the webometrics research community and a publicly available dataset manually classified by multiple coders should be made available to address the subjective nature of data sets.

More complex natural language processing techniques, for example, word sense disambiguation or other feature selection techniques like methods based on ant colony optimization (Aghdam, Ghasem-Aghaee and Basiri, 2009) or from the deviation from Poisson in text categorization (Ogura, Amano and Kondo, 2009) should also be investigated to assess whether they can be used to improve the accuracy of the classification models.

Although in this thesis, section 6.1 described a clustering technique that may be useful to policy makers, the results are based on an assumption that departments' websites describe all of their research interests, so it is necessary to carry out surveys and compare offline

similarity with the results of methods that were described in section 6.1. Text from computer science research groups were used for analysis, it is will be interesting to see the results when this technique is applied to other research disciplines.

The evidence from the results of the empirical studies shows that machine learning can improve the quality of web data for webometrics research. This opens the possibility the transfer of the machine learning techniques used in this thesis to the other webometrics analyses, not restricted to the academic web.

8. References

Aghdam, M. H., Ghasem-Aghaee, N. and Basiri, M. E. (2009) Text feature selection using ant colony optimization, *Expert Systems with Applications*, **36**(3), pp. 6843–6853, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0957417408005459 (Accessed 13 November 2014).

Aguillo, I. F., Granadino, B., Ortega, J. L. and Prieto, J. A. (2006) Scientific research activity and communication measured with cybermetrics indicators, *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, **57**(10), pp. 1296–1302, [online] Available from: http://dx.doi.org/10.1002/asi.20433.

Aguillo, I. F., Ortega, J. L. and Fernández, M. (2008) Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments, *Higher Education in Europe*, **33**(2-3), pp. 233 – 244, [online] Available from: http://dx.doi.org/10.1080/03797720802254031.

Almind, T. C. and Ingwersen, P. (1997) Informetric analyses on the world wide web: methodological approaches to 'webometrics,' *Jornal of Documentation*, **53**, p. 404; 404–426;

Amerijckx, C., Member, A., Verleysen, M., Thissen, P. and Legat, J. (1998) Image Compression by Self-Organized Kohonen Map, *IEEE transactions on neural networks*, **9**(3), pp. 503–507.

Amin, M. and Mabe, M. (2000) Impact factors: use and abuse, *Perspectives in publishing*, **1**(2), pp. 1–6, [online] Available from: https://info.aiaa.org/SC/PC/Private Documents/Journals Subcommittee Materials/IFUseandAbuse.pdf (Accessed 6 January 2014).

Aminpour, F., Kabiri, P., Otroj, Z. and Keshtkar, A. (2009) Webometric analysis of Iranian universities of medical sciences, *Scientometrics*, **80**(1), pp. 253–264, [online] Available from: http://dx.doi.org/10.1007/s11192-008-2059-y.

Anderson, J. D. (2006) *Qualitative and Quantitative Research*, Imperial County Office of Education, [online] Available from: http://www.icoe.org/webfm_send/1936.

Arakaki, M. and Willett, P. (2008) Webometric analysis of departments of librarianship and information science: a follow-up study, *Journal of Information Science*, **35**(2), pp. 143–152, [online] Available from: http://jis.sagepub.com/content/35/2/143.short (Accessed 8 July 2014).

Autant-Bernard, C., Billand, P., Frachisse, D. and Massard, N. (2007) Social distance versus spatial distance in R&D cooperation: Empirical evidence from European collaboration choices in micro and nanotechnologies, *Papers in Regional Science*, **86**(3), pp. 495–519, [online] Available from: http://doi.wiley.com/10.1111/j.1435-5957.2007.00132.x (Accessed 1 June 2014).

Ball, G. H. and Hall, D. J. (1965) *ISODATA. A novel method of data analysis and pattern classification*,.

Ballabio, D., Vasighi, M. and Filzmoser, P. (2013) Effects of supervised Self Organising Maps parameters on classification performance., *Analytica chimica acta*, Elsevier B.V., **765**, pp. 45–53, [online] Available from: http://www.ncbi.nlm.nih.gov/pubmed/23410625 (Accessed 6 June 2014).

Barabasi, A. L. and Albert, R. (1999) Emergence of scaling in random networks, *Science (New York, N.Y.)*, Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA., **286**(5439), pp. 509–512, [online] Available from: http://view.ncbi.nlm.nih.gov/pubmed/10521342.

Bar-Ilan, J. (2004) A microscopic link analysis of academic institutions within a country - the case of Israel, *Scientometrics*, AkadÃ©miai KiadÃ³, co-published with Springer Science+Business Media B.V., Formerly Kluwer Academic Publishers B.V, **59**(3), pp. 391–403, [online] Available from: http://dx.doi.org/10.1023/B:SCIE.0000018540.33706.c1.

Bar-Ilan, J. (2008) Informetrics at the beginning of the 21st century—A review, *Journal of Informetrics*, **2**(1), pp. 1–52, [online] Available from: http://www.sciencedirect.com/science/article/pii/S1751157707000740.

Bar-Ilan, J. (2005) What do we know about links and linking? A framework for studying links in academic environments, *Information Processing & Management*, **41**(4), pp. 973–986, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0306457304000135.

Bar-Ilan, J. (2002) Methods for measuring search engine performance over time, *Journal of the American Society for Information Science and Technology*, Wiley Online Library, **53**(4), pp. 308–319.

Bar-Ilan, J. and Peritz, B. C. (2004) Evolution, continuity, and disappearance of documents on a specific topic on the Web: A longitudinal study of "informetrics," *Journal of the American Society for Information Science and Technology*, Wiley Online Library, **55**(11), pp. 980–990.

Baroni, M. and Kilgarriff, A. (2006) Large linguistically-processed web corpora for multiple languages, In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations. Association for Computational Linguistics*, pp. 87–90, [online] Available from: http://dl.acm.org/citation.cfm?id=1608976 (Accessed 30 November 2014).

Beaver, D. D. (2001) Reflections on scientific collaboration (and its study): past, present, and future, *Scientometrics*, Akadémiai Kiadó,

co-published with Springer Science Business Media BV, Formerly Kluwer Academic Publishers BV, **52**(3), pp. 365–377.

Van den Besselaar, P. and Heimeriks, G. (2006) Mapping research topics using word-reference co-occurrences: A method and an exploratory case study, *Scientometrics*, Springer Netherlands, **68**(3), pp. 377–393, [online] Available from: http://dx.doi.org/10.1007/s11192-006-0118-9.

Bezdek, J. C. (1981) Pattern Recognition with Fuzzy Objective Function Algorithms, In Kluwer Academic Publishers.

Björneborn, L. (2006) 'Mini small worlds' of shortest link paths crossing domain boundaries in an academic Web space, *Scientometrics*, **68**(3), pp. 395–414, [online] Available from: http://link.springer.com/article/10.1007/s11192-006-0119-8 (Accessed 4 July 2014).

Björneborn, L. (2004) Small-world link structures across an academic web space: a library and information science approach, Det Informationsvidenskabelige AkademiDanish School of Library and Information Science, Institut østInstitut øst.

Björneborn, L. and Ingwersen, P. (2004) Toward a basic framework for webometrics, *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, **55**(14), pp. 1216–1227, [online] Available from: http://dx.doi.org/10.1002/asi.20077.

Blaxter, L., Hughes, C. and Tight, M. (1996) How to research, Open University Press (Buckingham and Bristol, PA, USA).

Boell, S. K., Wilson, C. S. and Cole, F. T. H. (2008) A webometric analysis of Australian Universities using staff and size dependent web impact factors (WIF), In Institute for Library and Information Science (IBI), p. 1.

Bra, P. De and Post, R. (1994) Information retrieval in the World-Wide Web: making client-based searching feasible, *Computer Networks and ISDN Systems*, **27**(2), pp. 183–192, [online] Available from: http://www.sciencedirect.com/science/article/pii/0169755294901325 (Accessed 22 November 2013).

Bradford, J., Kunz, C., Kohavi, R., Brunk, C. and Brodley, C. (1998) Pruning decision trees with misclassification costs, In *Machine Learning: ECML-98 SE - 18*, *Lecture Notes in Computer Science*, Nédellec, C. and Rouveirol, C. (eds.), Springer Berlin Heidelberg, pp. 131–136, [online] Available from: http://dx.doi.org/10.1007/BFb0026682.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984) *Classification and Regression Trees* , Wadsforth International Group.

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, **30**(1-7), pp. 107–117.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000) Graph structure in the Web, *Computer Networks*, **33**(1-6), pp. 309–320, [online] Available from: http://dx.doi.org/10.1016/S1389-1286(00)00083-9.

Burges, C. J. C. (1998) A Tutorial on Support Vector Machines for Pattern Recognition, *Data Minining and Knowledge Discovery*, Hingham, MA, USA, Kluwer Academic Publishers, **2**(2), pp. 121–167, [online] Available from: http://dx.doi.org/10.1023/A:1009715923555.

Castellano, G., Fanelli, A. M. and Pelillo, M. (1997) An iterative pruning algorithm for feedforward neural networks, *Neural Networks, IEEE Transactions on*, **8**(3), pp. 519–531.

Chakrabarti, S. (2003) *Mining the Web: Discovering knowledge from hypertext data*, Morgan Kaufmann, [online] Available from: http://books.google.co.uk/books?hl=en&lr=&id=5Zxw1h6yc_UC&oi=fnd&pg=PP2&dq=mining+the+web+chakrabarti&ots=ejVrAYJnwF&sig=OI5r1cQAzU4peolnKJUV94pQOsA (Accessed 22 November 2013).

Chakrabarti, S., Berg, M. Van den and Dom, B. (1999) Focused crawling: a new approach to topic-specific Web resource discovery, *Computer Networks*, **31**(11), pp. 1623–1640, [online] Available from: http://www.sciencedirect.com/science/article/pii/S1389128699000523 (Accessed 22 November 2013).

Chau, M. and Chen, H. (2008) A machine learning approach to web page filtering using content and structure analysis, *Decision Support Systems*, **44**(2), pp. 482–494, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0167923607000875.

Chen, Y., Tsai, F. S. and Chan, K. L. (2007) Blog search and mining in the business domain, In *Proceedings of the 2007 international workshop on Domain driven data mining - DDDM '07*, New York, New York, USA, ACM Press, pp. 55–60, [online] Available from: http://dl.acm.org/citation.cfm?id=1288552.1288560 (Accessed 30 May 2014).

Cho, J., Garcia-Molina, H. and Page, L. (1998) Efficient crawling through URL ordering, *Computer Networks and ISDN Systems*, **30**(1), pp. 161–172, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0169755298001081 (Accessed 22 November 2013).

Chu, H. (2005) Taxonomy of inlinked Web entities: What does it imply for webometric research?, *Library & Information Science Research*, **27**(1), pp. 8–27, [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0740818804000817 (Accessed 4 August 2014).

Cortes, C. and Vapnik, V. (1995) Support-Vector Networks, *Machine Learning*, **20**, p. 273; 273–297;

Cover, T. and Hart, P. (1967) Nearest neighbor pattern classification, *Information Theory, IEEE Transactions on*, **13**(1), pp. 21–27.

Cronin, B., Shaw, D. and La Barre, K. (2003) A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy, *Journal of the American Society for Information Science and Technology*, Wiley Online Library, **54**(9), pp. 855–871.

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A. and Callahan, E. (1998) Invoked on the Web, *Journal of the American Society for Information Science*, Wiley Subscription Services, Inc., A Wiley Company, **49**(14), pp. 1319–1328, [online] Available from: http://dx.doi.org/10.1002/(SICI)1097-4571(1998)49:14<1319::AID-ASI9>3.0.CO.

Cybermetrics, R. G. (2011) *The Academic Web Link Database Project*, Wolverhampton, University of Wolverhampton, [online] Available from: http://cybermetrics.wlv.ac.uk/database/.

Ding, C. and He, X. (2004) K -means clustering via principal component analysis, In *Twenty-first international conference on Machine learning - ICML '04*, New York, New York, USA, ACM Press, p. 29, [online] Available from: http://dl.acm.org/citation.cfm?id=1015330.1015408 (Accessed 11 June 2014).

Drineas, P., Frieze, A., Kannan, R., Vempala, S. and Vinay, V. (2004) Clustering Large Graphs via the Singular Value Decomposition, *Mach. Learn.*, **56**(1-3), pp. 9–33.

Dunn, J. C. (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics*, Taylor & Francis, **3**(3), pp. 32–57, [online] Available from: http://dx.doi.org/10.1080/01969727308546046.

Etzioni, O. (1996) The World-Wide Web: Quagmire or Gold Mine?, *Communications of the ACM*, New York, NY, USA, ACM, **39**(11), pp. 65–68, [online] Available from: http://doi.acm.org/10.1145/240455.240473.

Etzkowitz, H. and Leydesdorff, L. (1995) The Triple Helix--University-industry-government relations: A laboratory for knowledge based economic development, *Easst Review*, [online] Available from: http://dare.uva.nl/document/41280 (Accessed 24 July 2014).

Faba-Pérez, C., Zapico-Alonso, F., Guerrero-Bote, V. P. and Moya-Anegón, F. de (2005) Comparative analysis of webometric measurements in thematic environments, *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, **56**(8), pp. 779–785, [online] Available from: http://dx.doi.org/10.1002/asi.20161.

Figuerola, C. G. and Alonso Berrocal, J. L. (2013) Web link-based relationships among top European universities, *Journal of Information Science*, **39**(5), pp. 629–642, [online] Available from:

http://jis.sagepub.com/content/39/5/629.short (Accessed 15 June 2014).

François, C., Lamirel, J. and Shehabi, S. (2008) Combining advanced visualization and automatized reasoning for webometrics: a test study, *arXiv preprint arXiv:0810.5057*, [online] Available from: http://arxiv.org/abs/0810.5057 (Accessed 10 February 2014).

García-Aracil, A. and Palomares-Montero, D. (2009) Examining benchmark indicator systems for the evaluation of higher education institutions, *Higher Education*, **60**(2), pp. 217–234.

Gazni, A., Sugimoto, C. R. and Didegah, F. (2012) Mapping world scientific collaboration: authors, institutions, and countries, *Journal of the American Society for Information Science and Technology*, Wiley Online Library, **63**(2), pp. 323–335.

Giraudel, J. L. and Lek, S. (2001) A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination, *Ecological Modelling*, **146**(1–3), pp. 329–339, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0304380001003246.

Glanzel, W. and Schubert, A. (2005) Analysing Scientific Networks Through Co-Authorship, In *Handbook of Quantitative Science and Technology Research*, Moed, H. F., Glanzel, W., and Schmoch, U. (eds.), Springer Netherlands, pp. 257–276, [online] Available from: http://dx.doi.org/10.1007/1-4020-2755-9_12.

Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K. (2003) KNN Model-Based Approach in Classification, In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE SE - 62*, *Lecture Notes in Computer Science*, Meersman, R., Tari, Z., and Schmidt, D. (eds.), Springer Berlin Heidelberg, pp. 986–996, [online] Available from: http://dx.doi.org/10.1007/978-3-540-39964-3_62.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009) The WEKA data mining software: an update, *ACM SIGKDD Explorations Newsletter*, New York, NY, USA, ACM, **11**(1), pp. 10–18, [online] Available from: http://doi.acm.org/10.1145/1656274.1656278.

Hasan, K., Rahman, M. and Haque, A. (2009) Cost effective GPS-GPRS based object tracking system, In *Proceedings of the International MultiConference of Engineers and Computer Scientists,* pp. 18–20, [online] Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.148.8773&rep=rep1&type=pdf (Accessed 4 February 2014).

Hayfron-Acquah, J. and Gyimah, M. (2014) Classification and Recognition of Fingerprints using Self Organizing Maps (SOM)., *International Journal of Computer Science Issues*, **11**(1), pp. 153–159.

Heimeriks, G. and van den Besselaar, P. (2006) Analyzing hyperlinks networks: The meaning of hyperlink based indicators of knowledge

production., *Cybermetrics*, Isidro Aguillo, **10**(1), [online] Available from: http://cybermetrics.cindoc.csic.es/articles/v10i1p1.html (Accessed 15 June 2014).

Hérubel, J.-P. V. M. (1999) Historical bibliometrics: Its purpose and significance to the history of disciplines, JSTOR, pp. 380–388.

Heskes, T. (2001) Self-organizing maps, vector quantization, and mixture modeling., *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, **12**(6), pp. 1299–305, [online] Available from: http://www.ncbi.nlm.nih.gov/pubmed/18249959.

Highleyman, W. H. (1962) Linear Decision Functions, with Application to Pattern Recognition, *Proceedings of the IRE*, **50**(6), pp. 1501–1514.

Hoekman, J., Frenken, K. and Oort, F. (2008) The geography of collaborative knowledge production in Europe, *The Annals of Regional Science*, **43**(3), pp. 721–738, [online] Available from: http://link.springer.com/10.1007/s00168-008-0252-9 (Accessed 1 June 2014).

Hoekman, J., Frenken, K. and Tijssen, R. J. W. (2010) Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe, *Research Policy*, **39**(5), pp. 662–673, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0048733310000260 (Accessed 1 June 2014).

Hoerlesberger, M. and van den Besselaar, P. (2003) Mapping communication and collaboration in heterogeneous research networks , *Scientometrics*, **58**(2), pp. 391–413.

Holloway, T., Bozicevic, M. and Börner, K. (2007) Analyzing and visualizing the semantic coverage of Wikipedia and its authors, *Complexity*, **12**(3), pp. 30–40, [online] Available from: http://doi.wiley.com/10.1002/cplx.20164 (Accessed 3 July 2014).

Holmberg, K. (2009) Webometric Network Analysis: Mapping Cooperation and Geopolitical Connections Between Local Government Administration on the Web, {Å}bo Akademis F{ö}rlag, [online] Available from: http://books.google.co.uk/books?id=aL6JRQAACAAJ.

Holmberg, K. and Thelwall, M. (2014) Disciplinary differences in Twitter scholarly communication, *Scientometrics*, Springer Netherlands, pp. 1–16, [online] Available from: http://dx.doi.org/10.1007/s11192-014-1229-3.

Holmberg, K. and Thelwall, M. (2009) Local government web sites in Finland: A geographic and webometric analysis, *Scientometrics*, **79**(1), pp. 157 – 169, [online] Available from: http://dx.doi.org/10.1007/s11192-009-0410-6.

Huang, A. (2008) Similarity measures for text document clustering, In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand, pp. 49–56.

Ingwersen, P. (1998) The calculation of web impact factors , *Journal of Documentation*, Bingley, MCB UP Ltd, **54**(2; 54), pp. 236–243.

Ingwersen, P. and Björneborn, L. (2005) Methodological Issues of Webometric Studies, In *Handbook of Quantitative Science and Technology Research SE - 16*, Moed, H., Glänzel, W., and Schmoch, U. (eds.), Springer Netherlands, pp. 339–369, [online] Available from: http://dx.doi.org/10.1007/1-4020-2755-9_16.

Jacobs, D. (2010) Demystification of Bibliometrics, Scientometrics, Informetrics and Webometrics, In *11th DIS Annual Conference*, pp. 1–19.

Jaeger, R. M. (1990) *Statistics: A spectator sport*, Sage.

Jain, A. K. (2010) Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, **31**(8), pp. 651–666, [online] Available from: http://www.sciencedirect.com/science/article/pii/S016786550900323.

Jalal, S. K. (2010) Web impact factor and link analysis of selected Indian universities, Annals of Library and Information Studies, [online] Available from: https://drtc.isibang.ac.in//handle/1849/436 (Accessed 16 June 2014).

Janssens, F., Leta, J., Glänzel, W. and De Moor, B. (2006) Towards mapping library and information science, *Information Processing & Management*, **42**(6), pp. 1614–1642, [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S030645730600046X (Accessed 12 July 2014).

Jivani, A. (2011) A Comparative Study of Stemming Algorithms, *International Journal of Computer Technology and Applications*, **2**(6), pp. 1930–1938, [online] Available from: http://www.kenbenoit.net/courses/tcd2014qta/readings/Jivani_ijcta2011020632.pdf (Accessed 17 May 2014).

Katz, J. S. and Hicks, D. (1997) How much is a collaboration worth? A calibrated bibliometric model, *Scientometrics*, Springer, **40**(3), pp. 541–554.

Katz, J. S. and Martin, B. R. (1997) What is research collaboration?, *Research Policy*, **26**(1), pp. 1–18, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0048733396009171.

Kenekayoro, P., Buckley, K. and Thelwall, M. (2014a) Automatic classification of academic web page types, *Scientometrics*, Springer Netherlands, pp. 1–12, [online] Available from: http://dx.doi.org/10.1007/s11192-014-1292-9.

Kenekayoro, P., Buckley, K. and Thelwall, M. (2012) Fuzzy Clustering of UK Computer Science Departments, In *IADIS European Conference on Data Mining (DM)*, Ries, A. P. dos, Wang, P. S. P., and Abraham, A. P. (eds.), Lisbon, pp. 203–208.

Kenekayoro, P., Buckley, K. and Thelwall, M. (2014b) Hyperlinks as inter-university collaboration indicators, *Journal of Information Science*, **40**(4), pp. 514–522, [online] Available from: http://jis.sagepub.com/content/40/4/514 (Accessed 18 July 2014).

Kenekayoro, P., Buckley, K. and Thelwall, M. (2013) Motivation for Hyperlink Creation Using Inter-Page Relationships, In *14th Conference of the International Society for Scientometrics and. Informetrics (ISSI)*, Gorraiz, J., Schiebel, E., Gumpenberger, C., Hörlesberger, M., and Moed, H. (eds.), Vienna, pp. 1253–1269.

Khan, G. F. and Park, H. W. (2011) Measuring the triple helix on the web: Longitudinal trends in the university-industry-government relationship in Korea, *Journal of the American Society for Information Science and Technology*, **62**(12), pp. 2443–2455, [online] Available from: http://doi.wiley.com/10.1002/asi.21595 (Accessed 22 September 2014).

Kleinberg, J. M. (1999) Authoritative sources in a hyperlinked environment, *J. ACM*, **46**(5), pp. 604–632.

Kohlschütter, C., Fankhauser, P. and Nejdl, W. (2010) Boilerplate detection using shallow text features, In *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, New York, New York, USA, ACM Press, p. 441, [online] Available from: http://portal.acm.org/citation.cfm?doid=1718487.1718542.

Kohonen, T. (1990) The self-organizing map, *Proceedings of the IEEE*, **78**(9), pp. 1464–1480.

Kosala, R. and Blockeel, H. (2000) Web mining research: a survey, *Sigkdd Explorations Newsletter*, **2**(1), pp. 1–15.

Kotsiantis, S. B. (2013) Decision trees: a recent overview, *Artificial Intelligence Review*, Springer Netherlands, **30**(4), pp. 261–283, [online] Available from: http://dx.doi.org/10.1007/s10462-011-9272-4.

Kotsiantis, S. B. (2007) Supervised Machine Learning: A Review of Classification Techniques, In *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, Amsterdam, The Netherlands, The Netherlands, IOS Press, pp. 3–24, [online] Available from: http://dl.acm.org/citation.cfm?id=1566770.1566773.

Kotsiantis, S. B., Zaharakis, I. D. and Pintelas, P. E. (2006) Machine learning: a review of classification and combining techniques, *Artif.Intell.Rev.*, Norwell, MA, USA, Kluwer Academic Publishers, **26**(3), pp. 159–190, [online] Available from: http://dx.doi.org/10.1007/s10462-007-9052-3.

Kousha, K. and Thelwall, M. (2007) Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis, *Journal of the American Society for Information Science and Technology*, **58**(7), pp. 1055–1065, [online] Available from: http://onlinelibrary.wiley.com/doi/10.1002/asi.20584/full (Accessed 24 November 2013).

Kretschmer, H., Kretschmer, U. and Kretschmer, T. (2007) Reflection of co-authorship networks in the Web: Web hyperlinks versus Web visibility rates, *Scientometrics*, **70**(2), pp. 519–540, [online] Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-33846364551&partnerID=tZOtx3y1 (Accessed 15 June 2014).

Krippendorff, K. (2004) Reliability in Content Analysis., *Human Communication Research*, **30**(3), pp. 411–433, [online] Available from: http://doi.wiley.com/10.1111/j.1468-2958.2004.tb00738.x (Accessed 31 July 2014).

La, L., Guo, Q., Yang, D. and A Cao, Q. (2012) Multiclass Boosting with Adaptive Group-Based kNN and Its Application in Text Categorization, *Mathematical Problems in Engineering*, [online] Available from: http://dx.doi.org/10.1155/2012/793490.

Landry, R., Traore, N. and Godin, B. (1996) An econometric analysis of the effect of collaboration on academic research productivity, *Higher Education*, Springer, **32**(3), pp. 283–301.

Lau, K. W., Yin, H. and Hubbard, S. (2006) Kernel self-organising maps for classification, *Neurocomputing*, **69**(16-18), pp. 2033–2040, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0925231205003498 (Accessed 28 November 2014).

Lawrence, S. and Giles, C. L. (1999) Searching the web: General and scientific information access, In *Internet Technologies and Services, 1999. Proceedings. First IEEE/Popov Workshop on*, IEEE, pp. 18–31.

Lawrence, S. and Giles, C. L. (1998) Searching the world wide web, *Science*, American Association for the Advancement of Science, **280**(5360), pp. 98–100.

Lee, S. and Bozeman, B. (2005) The Impact of Research Collaboration on Scientific Productivity, *Social Studies of Science*, Sage Publications, Ltd., **35**(5, Scientific Collaboration), pp. 673–702, [online] Available from: http://www.jstor.org/stable/25046667.

Leydesdorff, L. and Etzkowitz, H. (1996) Emergence of a Triple Helix of university—industry—government relations, *Science and public policy*, [online] Available from: http://spp.oxfordjournals.org/content/23/5/279.short (Accessed 24 July 2014).

Leydesdorff, L. and Meyer, M. (2008) The triple helix model and the knowledge-based economy, *Scientometrics*, pp. 1–37, [online] Available from: http://bgarchives.bgu.ac.il/center/The Triple Helix Model and the Knowledge-Based Economy.pdf (Accessed 24 July 2014).

Leydesdorff, L. and Vaughan, L. (2006) Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment, *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, **57**(12), pp. 1616–1628, [online] Available from: http://dx.doi.org/10.1002/asi.20335.

Leydesdorff, L. and Welbers, K. (2011) The semantic mapping of words and co-words in contexts, *Journal of Informetrics*, Elsevier Ltd, **5**(3), pp. 469–475, [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S1751157711000095 (Accessed 28 May 2014).

Leydesdroff, L. (1989) Words and co-words as indicators of intellectual organization, *Research policy*, [online] Available from: http://www.sciencedirect.com/science/article/pii/0048733389900164 (Accessed 22 September 2014).

Liu, B. (2006) *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*, Secaucus, NJ, USA, Springer-Verlag New York, Inc.

Luo, P., Lin, F., Xiong, Y., Zhao, Y. and Shi, Z. (2009) Towards combining web classification and web information extraction: a case study, In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1235–1244.

Macqueen, J. B. (1967) Some methods of classification and analysis of multivariate observations, In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Marek, M., Pecina, P. and Spousta, M. (2007) Web Page Cleaning with Conditional Random Fields, In *Building and Exploring Web Corpora:*

Proceedings of the Fifth Web as Corpus Workshop, Incorporationg CleanEval (WAC3), pp. 155–162.

Marini, F., Zupan, J. and Magrì, A. L. (2004) On the use of counterpropagation artificial neural networks to characterize Italian rice varieties, *Analytica Chimica Acta*, **510**(2), pp. 231–240, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0003267004000479 (Accessed 6 June 2014).

Markovitch, S. and Rosenstein, D. (2002) Feature Generation Using General Constructor Functions, *Mach.Learn.*, Hingham, MA, USA, Kluwer Academic Publishers, **49**(1), pp. 59–98, [online] Available from: http://dx.doi.org/10.1023/A:1014046307775.

Mattsson, P., Laget, P., Vindefjärd, A. N. and Sundberg, C. J. (2010) What do European research collaboration networks in life sciences look like?, *Research Evaluation*, Oxford University Press, **19**(5), pp. 373–384.

Melin, G. and Persson, O. (1996) Studying research collaboration using co-authorships, *Scientometrics*, Springer, **36**(3), pp. 363–377.

Meloche, J. A. (2010) Visualization of the Chinese academic web based on social network analysis, *Journal of Information Science*, **36**(2), pp. 131–143, [online] Available from: http://jis.sagepub.com/content/36/2/131.short (Accessed 30 May 2014).

Menczer, F., Pant, G., Srinivasan, P. and Ruiz, M. (2001) Evaluating topic-driven web crawlers, In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 241–249, [online] Available from: http://dl.acm.org/citation.cfm?id=383995 (Accessed 22 November 2013).

Meyer, M. and Bhattacharya, S. (2004) Commonalities and differences between scholarly and technical collaboration. An exploration of co-invention and co-authorship analyses, *Scientometrics*, **61**(3), pp. 443–456.

Meyer, M., Siniläinen, T. and Utecht, J. (2003) Towards hybrid Triple Helix indicators: A study of university-related patents and a survey of academic inventors, *Scientometrics*, [online] Available from: http://www.akademiai.com/index/w5923u2714723g37.pdf (Accessed 24 July 2014).

Microsoft (2012) Top Keywords in Computer Science, [online] Available from: http://academic.research.microsoft.com/?SearchDomain=2&SubDomain=0&entitytype=8.

Minguillo, D. and Thelwall, M. (2012) Mapping the network structure of science parks: An exploratory study of cross-sectoral interactions reflected on the web, *Aslib Proceedings*, Emerald Group Publishing Limited, **64**(4), pp. 332–357, [online] Available from:

http://www.emeraldinsight.com/journals.htm?issn=0001-253X&volume=64&issue=4&articleid=17031115&show=abstract.

Minguillo, D. and Thelwall, M. (2011) The entrepreneurial role of the University: a link analysis of York Science Park, In *Proceedings of the ISSI 2011 Conference - 13th International Conference of the International Society for Scientometrics & Informetrics*, Noyons, E., Ngulube, P., and Leta, J. (eds.), pp. 570–583.

Mitchell, T. M. (2010) Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression, In *Machine Learning*, pp. 1–17.

Newman, M. E. J. (2004) Coauthorship Networks and Patterns of Scientific Collaboration, *Proceedings of the national academy of sciences*, **101**, pp. 5200–5205.

De Nooy, W., Mrvar, A. and Batagelj, V. (2005) *Exploratory social network analysis with Pajek*, Cambridge University Press.

Nwagwu, W. E. and Agarin, O. (2008) Nigerian University Websites: A Webometric Analysis, *Webology*, Webology, **5**(4), [online] Available from: http://www.webology.org/2008/v5n4/a65.html (Accessed 16 June 2014).

Ogura, H., Amano, H. and Kondo, M. (2009) Feature selection with a measure of deviations from Poisson in text categorization, *Expert Systems with Applications*, **36**(3), pp. 6826–6832, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0957417408005484 (Accessed 17 November 2014).

Olawoyin, R., Nieto, A., Grayson, R. L., Hardisty, F. and Oyewole, S. (2013) Application of artificial neural network (ANN)–self-organizing map (SOM) for the categorization of water, soil and sediment quality in petrochemical regions, *Expert Systems with Applications*, **40**(9), pp. 3634–3648, [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0957417412013103 (Accessed 31 May 2014).

Onyancha, O. B. and Ocholla, D. N. (2007) The Performance of South African and Kenyan universities on the world wide web : a web link analysis, *International Journal of Scientoemtrics Informetrics and Bibliometrics*, Cybermetrics, **11**(1), [online] Available from: http://uir.unisa.ac.za/handle/10500/5236 (Accessed 30 May 2014).

Ortega, J. and Aguillo, I. (2010a) Network collaboration in the 6th Framework Programmes: country participation in the health thematic area, *Scientometrics*, **84**(3), pp. 835–844, [online] Available from: http://www.akademiai.com/index/0642801m83128621.pdf (Accessed 7 January 2014).

Ortega, J. and Aguillo, I. (2010b) Shaping the European research collaboration in the 6th Framework Programme health thematic area through network analysis, *Scientometrics*, **85**(1), pp. 377–386, [online] Available from: http://www.akademiai.com/index/u206395t4076h042.pdf (Accessed 7 January 2014).

Ortega, J., Aguillo, I., Cothey, V. and Scharnhorst, A. (2008) Maps of the academic web in the European Higher Education Area â€" an exploration of visual web indicators, *Scientometrics*, **74**(2), pp. 295–308, [online] Available from: citeulike-article-id:2468532 http://dx.doi.org/10.1007/s11192-008-0218-9.

Ortega, J. L. and Aguillo, I. F. (2009) Mapping world-class universities on the web, *Information Processing & Management*, **45**(2), pp. 272–279, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0306457308001015 (Accessed 30 May 2014).

Ortega, J. L. and Aguillo, I. F. (2008) Visualization of the Nordic academic web: Link analysis using social network tools, *Information Processing & Management*, **44**(4), pp. 1624–1633, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0306457307001781 (Accessed 17 June 2014).

Ortega, J. L., Orduña-Malea, E. and Aguillo, I. F. (2013) Are web mentions accurate substitutes for inlinks for Spanish universities?, *Online Information Review*, Emerald Group Publishing Limited, **38**(1), pp. 59–77, [online] Available from: http://www.emeraldinsight.com/journals.htm?issn=1468-4527&volume=38&issue=1&articleid=17102931&show=html (Accessed 15 June 2014).

Ozel, B. and Park, H. W. (2011) Online image content analysis of political figures: an exploratory study, *Quality & Quantity*, **46**(4), pp. 1013–1024, [online] Available from: http://link.springer.com/10.1007/s11135-011-9445-x (Accessed 12 August 2014).

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999) The PageRank citation ranking: bringing order to the web., Stanford InfoLab.

Pant, G., Srinivasan, P. and Menczer, F. (2004) Crawling the Web, In *In Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Edited by M. Levene and A. Poulovassilis*, Springer-Verlag, pp. 153–178.

Parekh, R., Yang, J. and Honavar, V. (2000) Constructive neural-network learning algorithms for pattern classification, *Neural Networks, IEEE Transactions on*, **11**(2), pp. 436–451.

Park, H. W. and Thelwall, M. (2008) Link analysis: Hyperlink patterns and social structure on politicians' Web sites in South Korea, *Quality & Quantity*, **42**(5), pp. 687 – 697, [online] Available from: http://dx.doi.org/10.1007/s11135-007-9109-z.

Payne, N. and Thelwall, M. (2004) A Statistical Analysis of UK Academic Web Links, *International Journal of Scientometrics, Informetrics and Bibliometrics*, **8**(1).

Perianes-Rodríguez, A., Olmeda-Gómez, C. and Moya-Anegón, F. (2009) Detecting, identifying and visualizing research groups in co-authorship networks, *Scientometrics*, **82**(2), pp. 307–319, [online]

Available from: http://link.springer.com/10.1007/s11192-009-0040-z (Accessed 7 October 2014).

Peters, H. P. F. and van Raan, a. F. J. (1993) Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling, *Research Policy*, **22**(1), pp. 23–45, [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/004873339390031C.

Ponds, R., van Oort, F. and Frenken, K. (2007) The geographical and institutional proximity of research collaboration, *Papers in Regional Science*, **86**(3), pp. 423–443, [online] Available from: http://doi.wiley.com/10.1111/j.1435-5957.2007.00126.x (Accessed 1 June 2014).

Potratz, W. and Widmaier, B. (1996) Industrial transformation in central and eastern europe: Is innovation a way to integration?, *MOCT-MOST: Economic Policy in Transitional Economies*, **6**(4), pp. 55–70.

Priego, J. (2003) A methodological approach to the Triple Helix dimensionality: A comparative study of Biology and Biomedicine Centres of two European National Research Councils, *Scientometrics*, **58**(2), pp. 429–443, [online] Available from: http://www.akademiai.com/index/x310u18116712808.pdf (Accessed 24 July 2014).

Qi, X. and Davison, B. D. (2009) Web page classification: Features and algorithms, *ACM Computing Surveys (CSUR)*, **41**(2), pp. 1–31.

Quinlan, J. R. (1993) *C4.5: programs for machine learning*, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

Rafols, I. and Meyer, M. (2010) Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience, *Scientometrics*, pp. 1–28, [online] Available from: http://www.akademiai.com/index/h75t85p933471v27.pdf (Accessed 24 July 2014).

Research Council UK (2013) Gateway to Research, [online] Available from: http://www.gtr.rcuk.ac.uk/ (Accessed 20 May 2013).

Roediger-Schluga, T. and Barber, M. (2008) R&D collaboration networks in the European Framework Programmes: Data processing, network construction and selected results, *International Journal of Foresight and Innovation Policy*, **4**(3), pp. 321–347, [online] Available from: http://inderscience.metapress.com/index/h1337472726u163l.pdf (Accessed 7 January 2014).

Romero-Frías, E. and Vaughan, L. (2012) Exploring the relationships between media and political parties through web hyperlink analysis: The case of Spain, *Journal of the American Society for Information Science and Technology*, **63**(5), pp. 967–976, [online] Available from: http://dx.doi.org/10.1002/asi.22625.

Rosenblatt, F. (1962) *Principles of neurodynamics; perceptrons and the theory of brain mechanisms*, Washington, Spartan Books.

Rousseau, R. (1997) Sitations: an exploratory study, *International Journal of Scientometrics, Informetrics and Bibliometrics*, **1**(1), [online] Available from: http://www.webcitation.org/5stBoPIrC.

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E. and Suter, B. W. (1990) The multilayer perceptron as an approximation to a Bayes optimal discriminant function., *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, **1**(4), pp. 296–8, [online] Available from: http://www.ncbi.nlm.nih.gov/pubmed/18282850.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) Learning representations by back-propagating errors, *Nature*, **323**, pp. 533–536, [online] Available from: citeulike-article-id:6136933.

Schaeffer, S. E. (2007) Graph clustering, *Computer Science Review*, **1**(1), pp. 27–64, [online] Available from: http://www.sciencedirect.com/science/article/pii/S1574013707000020 (Accessed 26 May 2014).

Scharnhorst, A. and Wouters, P. (2006) Webindicators: a new generation of S&T indicators, *Cybermetrics*, **10**(1), [online] Available from: http://depot.knaw.nl/3752 (Accessed 4 July 2014).

Schreiber, M., Malesios, C. C. and Psarakis, S. (2012) Exploratory factor analysis for the Hirsch index, 17 h-type variants, and some traditional bibliometric indicators, *Journal of Informetrics*, **6**(3), pp. 347–358, [online] Available from:

http://www.sciencedirect.com/science/article/pii/S1751157712000107 (Accessed 7 October 2014).

Seeber, M., Lepori, B., Lomi, A., Aguillo, I. and Barberio, V. (2012) Factors affecting web links between European higher education institutions, *Journal of Informetrics*, **6**(3), pp. 435–447, [online] Available from: http://www.sciencedirect.com/science/article/pii/S1751157712000235 (Accessed 23 May 2014).

Shari, S., Haddow, G. and Genoni, P. (2012) Bibliometric and webometric methods for assessing research collaboration, *Library Review*, Emerald Group Publishing Limited, **61**(8), pp. 592–607, [online] Available from: http://www.emeraldinsight.com/journals.htm?issn=0024-2535&volume=61&issue=8/9&articleid=17065497&show=html (Accessed 15 June 2014).

Shef.ac.uk (2014) Funding of research in UK Higher Education - Higher Education - How finance works - Information for Staff - Finance - The University of Sheffield, [online] Available from: http://www.shef.ac.uk/finance/staff-information/howfinanceworks/higher_education/funding_of_research (Accessed 6 January 2014).

Singh, S. K., Paini, D. R., Ash, G. J. and Hodda, M. (2013) Prioritising plant-parasitic nematode species biosecurity risks using self organising maps, *Biological Invasions*, [online] Available from:

http://link.springer.com/10.1007/s10530-013-0588-7 (Accessed 5 June 2014).

Smith, A. (1999) A tale of two Web spaces: Comparing sites using Web impact factors , *Journal of Documentation*, CiteULike.org, **55**(5), pp. 577–592, [online] Available from: http://www.citeulike.org/user/dreymond33/article/6091839.

Smith, A. (2003a) Classifying links for substantive Web Impact Factors, In *Proceedings of the 9th International Conference on Scientometrics and Informetrics*.

Smith, A. (2003b) Think local, search global? Comparing search engines for searching geographically specific information, *Online Information Review*, MCB UP Ltd, **27**(2), pp. 102–109.

Sonnenwald, D. (2007) Scientific collaboration, *Annual review of information science and technology*, **41**(1), pp. 643–681, [online] Available from: http://onlinelibrary.wiley.com/doi/10.1002/aris.2007.1440410121/full (Accessed 30 January 2014).

Stuart, D. (2008) Web Manifestations of Knowledge-based Innovation Systems in the UK, [online] Available from: http://wlv.openrepository.com/wlv/handle/2436/33737 (Accessed 6 January 2014).

Stuart, D., Thelwall, M. and Harries, G. (2007) UK academic web links and collaboration - an exploratory study, *Journal of Information*

Science, **33**(2), pp. 231 – 246, [online] Available from: http://dx.doi.org/10.1177/0165551506075326.

Subramanyam, K. (1983) Bibliometric studies of research collaboration: A review, *Journal of Information Science*, Sage Publications, **6**(1), pp. 33–38.

Sun, Y. (2000) On quantization error of self-organizing map network, *Neurocomputing*, **34**, pp. 169–193, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0925231200002927 (Accessed 6 June 2014).

Tan, P.-N., Steinbach, M. and Kumar, V. (2005) *Introduction to Data Mining,* Boston, MA, USA, Addison-Wesley Longman Publishing Co., Inc.

Thelwall, M. (2002a) A research and institutional size-based model for national university Web site interlinking, *Journal of Documentation*, **58**(6), pp. 683–694, [online] Available from: http://www.emeraldinsight.com/journals.htm?articleid=864204&show=abstract (Accessed 16 January 2014).

Thelwall, M. (2002b) An initial exploration of the link relationship between UK university Web sites., *ASLIB Proceedings*, **52**(2), pp. 118–126, [online] Available from: http://www.emeraldinsight.com/10.1108/00012530210435248.

Thelwall, M. (2002c) Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university Web sites, *Journal of the American Society for*

Information Science and Technology, Wiley Subscription Services, Inc., A Wiley Company, **53**(12), pp. 995–1005, [online] Available from: http://dx.doi.org/10.1002/asi.10135.

Thelwall, M. (2002d) Evidence for the existence of geographic trends in university Web site interlinking, *Journal of Documentation*, **58**(5), pp. 563–574.

Thelwall, M. (2001a) Exploring the link structure of the Web with network diagrams, *Journal of Information Science*, Sage Publications, **27**(6), pp. 393–401.

Thelwall, M. (2008a) Extracting accurate and complete results from search engines: case study Windows Live, *Journal of the American Society for Information Science and Technology*, Wiley Online Library, **59**(1), pp. 38–50.

Thelwall, M. (2001b) Extracting macroscopic information from Web links, *Journal of the American Society for Information Science and Technology*, John Wiley & Sons, Inc., **52**(13), pp. 1157–1168, [online] Available from: http://dx.doi.org/10.1002/asi.1182.

Thelwall, M. (2006) Interpreting social science link analysis research: A theoretical framework, *Journal of the American Society for Information Science*, New York, NY, USA, John Wiley \& Sons, Inc, **57**(1), pp. 60–68, [online] Available from: http://dx.doi.org/10.1002/asi.v57:1.

Thelwall, M. (2009) Introduction to Webometrics: Quantitative Web Research for the Social Sciences, *Synthesis Lectures on Information*

Concepts, Retrieval, and Services, Morgan & Claypool Publishers, **1**(1), pp. 1–116, [online] Available from: http://dx.doi.org/10.2200/S00176ED1V01Y200903ICR004.

Thelwall, M. (2004) *Link analysis: An information science approach*, [online] Available from: http://www.citeulike.org/group/1840/article/1191506 (Accessed 6 January 2014).

Thelwall, M. (2008b) Quantitative comparisons of search engine results, *Journal of the American Society for Information Science and Technology*, Wiley Online Library, **59**(11), pp. 1702–1710.

Thelwall, M. (2002e) The top 100 linked-to pages on UK university web sites: high inlink counts are not usually associated with quality scholarly content, *Journal of Information Science*, **28**(6), pp. 483 – 491, [online] Available from: http://dx.doi.org/10.1177/016555150202800604.

Thelwall, M. (2003) What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation., *Information Research*, **8**(3), [online] Available from: http://informationr.net/ir/8-3/paper151.html.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010) Sentiment strength detection in short informal text, *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, **61**(12), pp. 2544–2558, [online] Available from: http://dx.doi.org/10.1002/asi.21416.

Thelwall, M. and Harries, G. (2003) The connection between the research of a university and counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university, *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, **54**(7), pp. 594–602, [online] Available from: http://dx.doi.org/10.1002/asi.10161.

Thelwall, M. and Hasler, L. (2007) Blog search engines, *Online Information Review*, Emerald Group Publishing Limited, **31**(4), pp. 467–479, [online] Available from: http://www.emeraldinsight.com/journals.htm?issn=1468-4527&volume=31&issue=4&articleid=1621795&show=html (Accessed 30 May 2014).

Thelwall, M., Klitkou, A., Verbeek, A., Stuart, D. and Vincent, C. (2010) Policy-relevant Webometrics for individual scientific fields, *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, **61**(7), pp. 1464–1475, [online] Available from: http://dx.doi.org/10.1002/asi.21345.

Thelwall, M. and Price, L. (2003) Disciplinary differences in academic web presence–a statistical study of the UK, *Libri*, **53**, pp. 242–253, [online] Available from: http://www.degruyter.com/view/j/libr.2003.53.issue-4/libr.2003.242/libr.2003.242.xml (Accessed 2 June 2014).

Thelwall, M. and Stuart, D. (2006) Web crawling ethics revisited: Cost, privacy, and denial of service, *Journal of the American Society for Information Science and Technology*, **57**(13), pp. 1771–1779.

Thelwall, M. and Sud, P. (2011) A comparison of methods for collecting web citation data for academic organizations, *Journal of the American Society for Information Science and Technology*, Wiley Online Library, **62**(8), pp. 1488–1497.

Thelwall, M. and Sud, P. (2012) Webometric research with the Bing Search API 2.0, *Journal of Informetrics*, Elsevier, **6**(1), pp. 44–52.

Thelwall, M., Sud, P. and Wilkinson, D. (2012) Link and co-inlink network diagrams with URL citations or title mentions, *Journal of the American Society for Information Science and Technology*, **63**(4), pp. 805–816, [online] Available from: http://dx.doi.org/10.1002/asi.21709.

Thelwall, M., Vann, K. and Fairclough, R. (2006) Web issue analysis: An integrated water resource management case study, *Journal of the American Society for Information Science and Technology*, Wiley Online Library, **57**(10), pp. 1303–1314.

Thelwall, M., Vaughan, L., Cothey, V., Li, X. and Smith, A. G. (2003) Which academic subjects have most online impact? A pilot study and a new classification process, *Online Information Review*, **27**(5), pp. 333–343, [online] Available from: http://www.emeraldinsight.com/10.1108/14684520310502298.

Thelwall, M. and Wilkinson, D. (2008) A generic lexical URL segmentation framework for counting links, colinks or URLs, *Library & Information Science Research*, **30**(2), pp. 94–101, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0740818808000315.

Thelwall, M. and Wilkinson, D. (2003) Graph structure in three national academic Webs: Power laws with anomalies, *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, **54**(8), pp. 706–712, [online] Available from: http://dx.doi.org/10.1002/asi.10267.

Thelwall, M. and Zuccala, A. (2008) A university-centred European Union link analysis, *Scientometrics*, **45**(3), pp. 407–420.

Thijs, B. and Glänzel, W. (2010) A structural analysis of collaboration between European research institutes, *Research Evaluation*, Oxford University Press, **19**(1), pp. 55–65.

Thomas, O. and Willett, P. (2000) Webometric analysis of departments of librarianship and information science, *Journal of Information Science*, **26**(6), pp. 421–428, [online] Available from: http://jis.sagepub.com/content/26/6/421.short (Accessed 8 July 2014).

Tuomaala, O., Järvelin, K. and Vakkari, P. (2014) Evolution of library and information science, 1965-2005: Content analysis of journal articles, *Journal of the Association for Information Science and*

Technology, **65**(7), pp. 1446–1462, [online] Available from: http://doi.wiley.com/10.1002/asi.23034 (Accessed 29 July 2014).

Utgoff, P. (1989) Incremental Induction of Decision Trees, *Machine Learning*, Kluwer Academic Publishers-Plenum Publishers, **4**(2), pp. 161–186, [online] Available from: http://dx.doi.org/10.1023/A:1022699900025.

Vaseleiadou, E. and van den Besselaar, P. (2006) Linking Shallow, linking deep. How scientific intermediaries use the web for their network of collaborators, *Internationl Journal of Scientometrics, Informetrics and Bibliometrics*, **10**(1).

Vaughan, L. (2005) Mining Web hyperlink data for business information: The case of telecommunications equipment companies, In *Proceedings of The 1st International Conference on Signal-Image Technology & Internet-Based Systems*, p. 190; 190–195;

Vaughan, L. and Gao, Y. (2006) Why are hyperlinks to business Websites created? A content analysis , *Scientometrics*, **67**, p. 291; 291–300; –300.

Vaughan, L., Kipp, M. and Gao, Y. (2007) Why are Websites co-linked? The case of Canadian universities , *Scientometrics*, Springer, **72**(1. (July 2007), pp. 81-92, doi), pp. 10–92, [online] Available from: http://www.citeulike.org/user/meikipp/article/1602477.

Vaughan, L. and Romero-Frías, E. (2012) Exploring Web keyword analysis as an alternative to link analysis: a multi-industry case,

Scientometrics, Springer Netherlands, **93**(1), pp. 217–232, [online] Available from: http://dx.doi.org/10.1007/s11192-012-0640-x.

Vaughan, L. and Romero-Frías, E. (2010) Web hyperlink patterns and the financial variables of the global banking industry, *Journal of Information Science* , **36** (4), pp. 530–541, [online] Available from: http://jis.sagepub.com/content/36/4/530.abstract.

Vaughan, L., Tang, J. and Du, J. (2009) Examining the robustness of web co-link analysis, *Online Information Review*, Emerald Group Publishing Limited, **33**(5), pp. 956–972.

Vaughan, L. and Thelwall, M. (2005) A modeling approach to uncover hyperlink patterns: the case of Canadian universities, *Information Processing & Management*, **41**(2), pp. 347–359, [online] Available from: http://www.sciencedirect.com/science/article/pii/S0306457303000840.

Vaughan, L. and Thelwall, M. (2003) Scholarly use of the Web: What are the key inducers of links to journal Web sites?, *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, **54**(1), pp. 29–38, [online] Available from: http://dx.doi.org/10.1002/asi.10184.

Vaughan, L. and Wu, G. (2004) Links to commercial websites as a source of business information, *Scientometrics*, **60**(3), pp. 487 – 496, [online] Available from: http://dx.doi.org/10.1023/b:scie.0000034389.14825.bc.

Vaughan, L. and You, J. (2008) Content assisted web co-link analysis for competitive intelligence, *Scientometrics*, Akadémiai Kiadó, co-published with Springer Science Business Media BV, Formerly Kluwer Academic Publishers BV, **77**(3), pp. 433–444.

Vaughan, L. and You, J. (2009) Keyword Enhanced Web Structure Mining for Business Intelligence, *Advanced Internet Based Systems and Applications*, **4879**, pp. 161 – 168, [online] Available from: http://dx.doi.org/10.1007/978-3-642-01350-8_15.

Vaughan, L. and You, J. (2010) Word co-occurrences on Webpages as a measure of the relatedness of organizations: A new Webometrics concept, *Journal of Informetrics*, **4**(4), pp. 483 – 491, [online] Available from: http://dx.doi.org/10.1016/j.joi.2010.04.005.

Vesanto, J. (1999) SOM-based data visualization methods, *Intelligent Data Analysis*, **3**(2), pp. 111–126, [online] Available from: http://dx.doi.org/10.3233/IDA-1999-3203.

Vesanto, J. and Alhoniemi, E. (2000) Clustering of the self-organizing map, *IEEE transactions on neural networks*, **11**(3), pp. 586–600, [online] Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=846731 (Accessed 28 November 2014).

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. (1999) Self-organizing map in Matlab: the SOM Toolbox, In *Proceedings of the Matlab DSP conference*, pp. 16–17, [online] Available from:

http://cda.psych.uiuc.edu/matlab_class/martinez/edatoolbox/Docs/toolbox2paper.pdf (Accessed 3 March 2014).

Weller, M. (2011) *The Digital Scholar: How Technology is Transforming Academic Practice*, Bloomsbury Academic, p. 55, [online] Available from: http://books.google.co.uk/books?id=KV1MAQAAQBAJ.

Weninger, T. (2012) Web Sites as Schemas - A tree matching Approach, In *Command, Control and Interoperability Center for Advanced Data Analysis*, Illinois.

Weninger, T., Johnston, T. J. and Han, J. (2013) The parallel path framework for entity discovery on the web, *ACM Transactions on the Web*, ACM, **7**(3), pp. 1–29, [online] Available from: http://dl.acm.org/citation.cfm?id=2516633.2516638 (Accessed 30 May 2014).

Whittaker, J., Courtial, J., Law, J. and Whittakert, J. (1989) Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis, *Social Studies of Science*, **19**(3), pp. 473–496.

Wilkinson, D., Harries, G., Thelwall, M. and Price, L. (2003) Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication, *Journal of Information Science*, **29**(1), pp. 49–56, [online] Available from: http://jis.sagepub.com/content/29/1/49.abstract.

Wolpert, D. H. and Macready, W. G. (1997) No free lunch theorems for optimization, *IEEE Transactions on Evolutionary Computation*, IEEE, **1**(1), pp. 67–82, [online] Available from:

http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=585893 (Accessed 27 May 2014).

Wright, S. (1921) Correlation and causation, *Journal of agricultural research*, [online] Available from: http://www.ssc.wisc.edu/soc/class/soc952/Wright/Wright_Correlation and Causation.pdf (Accessed 13 June 2014).

Xuemei, L., Thelwall, M., Musgrove, P. and Wilkinson, D. (2003) The relationship between the WIFs or inlinks of Computer Science Departments in UK and their RAE ratings or research productivities in 2001 , *Scientometrics*, **57**(2), pp. 239–255.

Yin, H. (2008) The Self-Organizing Maps: Background, Theories, Extensions and Applications, In *Computational Intelligence: A Compendium SE - 17*, *Studies in Computational Intelligence*, Fulcher, J. and Jain, L. C. (eds.), Springer Berlin Heidelberg, pp. 715–762, [online] Available from: http://dx.doi.org/10.1007/978-3-540-78293-3_17.

Yu, L. and Liu, H. (2004) Efficient Feature Selection via Analysis of Relevance and Redundancy, *Journal of Machine Learning Research*, JMLR.org, **5**, pp. 1205–1224, [online] Available from: http://dl.acm.org/citation.cfm?id=1005332.1044700.

Zhang, Q. and Segall, R. S. (2008) Web Mining: A Survey of Current Research, Techniques, and Software, *International Journal of Information Technology & Decision Making*, **07**(04), p. 683, [online] Available from: http://dx.doi.org/10.1142/s0219622008003150.

Zhao, D. and Strotmann, A. (2008) Information science during the first decade of the web: An enriched author cocitation analysis, *Journal of the American Society for Information Science and Technology*, **59**(6), pp. 916–937, [online] Available from: http://doi.wiley.com/10.1002/asi.20799 (Accessed 24 September 2014).

Zhou, Z. and Zhang, M. (2006) Multi-Instance Multi-Label Learning with Application to Scene Classification, In *Advances in Neural Information Processing Systems*, Schölkopf, B., Platt, J., and Hoffman, T. (eds.), MIT Press, pp. 1609–1616, [online] Available from: http://books.nips.cc/papers/files/nips19/NIPS2006_0348.pdf.

9. Appendices

Appendix A: List of UK universities that are used in Chapter 5.

Anglia Ruskin University	Open University	University of Hull
Aberystwyth University	Oxford Brookes University	University of Kent
Aston University	Plymouth University	University of Leeds
Bangor University	Queen Margaret University Edinburgh	University of Leicester
Bath Spa University	Queen Mary University London	University of Lincoln
Birkbeck, University of London	Queens' University Belfast	University of Liverpool
Birmingham City University	Robert Gordon university	University of Manchester
Bournemouth University	Royal Holloway university	University of Northampton
Bristol University	School of Oriental and African Studies	University of Northumbria
Brunel University	Sheffield Hallam University	University of Nottingham
Canterbury Christ Church University	Swansea University	University of Oxford
City University London	Teeside University	University of Portsmouth
Coventry University	University College London	University of Reading
De Montfort University	University for the Creative Arts	University of Salford
Durham University	University of Aberdeen	University of Sheffield
Edinburgh Napier University	University of Abertay	University of Southampton
Glasgow Caledonian University	University of Bath	University of St Andrews
Goldsmiths, University of London	University of Bedfordshire	University of Stirling
Harper Adams University	University of Birmingham	University of Strathclyde
Heriot-Watt University	University of Bradford	University of Sunderland
Imperial College London	University of Brighton	University of Surrey
Kings College London	University of Buckingham	University of Sussex
Kingston University	University of Cambridge	University of the Arts London
Lancaster University	University of Derby	University of the West of England

Leeds Metropolitan University	University of Dundee	University of Ulster
Liverpool Hope University	University of East Anglia	University of Wales, Lampeter
Liverpool John Moores University	University of East London	University of Wales, Newport
London Metropolitan University	University of Edinburgh	University of Warwick
London Sch of Economics and Political Science	University of Exeter	University of West London
London Sch of Hygiene and Trop Med	University of Glamorgan	University of West of Scotland
London South Bank University	University of Glasgow	University of Westminster
Loughborough University	University of Gloucestershire	University of Wolverhampton
Manchester Metropolitan University	University of Greenwich	University of Worcester
Newcastle University	University of Hertfordshire	University of York
Nottingham Trent University	University of Huddersfield	

Appendix B: List of Computer Science Research Groups used in Chapter 6.

Research Group	**URL**
Agents Reasoning Knowledge	http://www.abdn.ac.uk/ncs/computing/research/ark/
Natural Language Generation	http://www.abdn.ac.uk/ncs/computing/research/nlg/
Systems Modelling	http://www.abdn.ac.uk/ncs/computing/research/sysmod/
Modelling And Visualization	http://www.abertay.ac.uk/research/modelling/
Humans And Technology	http://www.abertay.ac.uk/research/tech/
Advanced Reasoning Group	http://www.aber.ac.uk/en/cs/research/ar/
Bioinformatics	http://www.aber.ac.uk/en/cs/research/cb/
Intelligent Robotics	http://www.aber.ac.uk/en/cs/research/ir/
Visualization And Graphics	http://www.aber.ac.uk/en/cs/research/vgv/
Digital Modelling	http://www.anglia.ac.uk/ruskin/en/home/microsites/dmrg.html
Sound And Audio Engineering	http://www.anglia.ac.uk/ruskin/en/home/faculties/fst/departments/comptech/research/technology/research.html

Engineering Analysis	http://www.anglia.ac.uk/ruskin/en/home/faculties/fst/departments/comptech/research/engineering/simulation.html
Telecommunications	http://www.anglia.ac.uk/ruskin/en/home/faculties/fst/departments/comptech/research/telecommunications.html
Computer Science Research Group	http://www1.aston.ac.uk/eas/research/groups/csrg/research-areas/
Visualization And Graphics	http://www.vmg.cs.bangor.ac.uk/
Systems And Modeling	http://www.bangor.ac.uk/cs/research/systems.php
Pattern Recognition And Machine Learning	http://www.bangor.ac.uk/cs/research/prml2.php
Artificial Intelligence And Intelligent Agents	http://www.bangor.ac.uk/cs/research/aiia2.php
Human Computer Interaction	http://www.bath.ac.uk/comp-sci/research/hci.html
Mathematical Foundations	http://www.bath.ac.uk/comp-sci/research/mathsfoundations.html
Media Technology	http://www.bath.ac.uk/comp-sci/research/mediatechnology.html
Artificial Intelligence	http://www.bath.ac.uk/comp-sci/research/ai.html
Computer Graphics And Visualization	http://www.beds.ac.uk/research/irac/ccgv
Distributed Technology	http://www.beds.ac.uk/research/irac/credit
Wireless Technology	http://www.beds.ac.uk/research/irac/cwr
Natural Computation	http://www.cs.bham.ac.uk/research/groupings/natural/computation/
Machine Learning	http://www.cs.bham.ac.uk/research/groupings/machine/learning/
Robotics	http://www.cs.bham.ac.uk/research/groupings/robotics/and/cognitive/architectures/overview/
Medical Image Interpretation	http://www.cs.bham.ac.uk/research/groupings/medical/image/interpretation/
Reasoning	http://www.cs.bham.ac.uk/research/groupings/reasoning/
Advanced Interaction	http://www.cs.bham.ac.uk/research/groupings/language/and/interaction/human/computer/interaction/
Natural Language Processing	http://www.cs.bham.ac.uk/research/groupings/language/and/interaction/natural/language/processing/
Scientific Document Analysis	http://www.cs.bham.ac.uk/research/groupings/reasoning/sdag/
Formal Verification And Security	http://www.cs.bham.ac.uk/research/groupings/formal/verification/and/security/
Parallel And Distributed Computing	http://www.cs.bham.ac.uk/research/groupings/parallel/and/distributed/computing/
Mathematics Foundation	http://www.cs.bham.ac.uk/research/groupings/mathematics/foundations/

Principles Of Programming	http://www.cs.bham.ac.uk/research/groupings/principles/of/programming/
Software Engineering	http://www.cs.bham.ac.uk/research/groupings/software/engineering/
City Research	http://www.bcu.ac.uk/tee/ctn/research
Creative Research	http://dec.bournemouth.ac.uk/research/creative/about.html
Smart Technology	http://www.bournemouth.ac.uk/strc/themes.html
Software Systems	http://www.bournemouth.ac.uk/ssrc/themes.html
Artificial Intelligence	http://www.brad.ac.uk/research/research-in-schools/school-of-computing-informatics-and-media/artificial-intelligence/
Applied Mathematics	http://www.brad.ac.uk/research/research-in-schools/school-of-computing-informatics-and-media/applied-mathematics/
Digital Imaging	http://www.brad.ac.uk/research/research-in-schools/school-of-computing-informatics-and-media/digital-imaging-and-visualisation/
Networks	http://www.brad.ac.uk/research/research-in-schools/school-of-computing-informatics-and-media/networks-and-performance-engineering/
Visual Computing	http://www.brad.ac.uk/research/research-in-schools/school-of-computing-informatics-and-media/Centre-for-visual-computing/
Computational Intelligence	http://www.cem.brighton.ac.uk/research/cig/
Visual Modelling	http://www.cem.brighton.ac.uk/research/vmg/
Natural Language	http://www.nltg.brighton.ac.uk/nltg/
Computer Vision	http://www.cs.bris.ac.uk/Research/Vision/
Cryptography	http://www.cs.bris.ac.uk/Research/CryptographySecurity/
Microelectronics	http://www.cs.bris.ac.uk/Research/Micro/
Intelligent Systems	http://intelligentsystems.bristol.ac.uk/
Interaction And Graphics	http://big.cs.bris.ac.uk/
Information And Knowledge Management	http://www.brunel.ac.uk/siscm/disc/research/cikm
Information Systems	http://www.brunel.ac.uk/siscm/disc/research/cisr
Intelligent Data Analysis	http://www.brunel.ac.uk/siscm/disc/research/cida
People And Interactivity	http://www.brunel.ac.uk/siscm/disc/research/pandi
Wireless Network Group	http://www.buckingham.ac.uk/sciences/dept/appliedcomputing/wirelessnetworkgroup
Applied Computing	http://www.buckingham.ac.uk/research/appliedcomputing
Artificial Intelligence	http://www.cl.cam.ac.uk/research/ai/
Computer Architecture	http://www.cl.cam.ac.uk/research/comparch/
Digital Technology	http://www.cl.cam.ac.uk/research/dtg/www/
Graphics And	http://www.cl.cam.ac.uk/research/rainbow/research/

Interaction	
Natural Language	http://www.cl.cam.ac.uk/research/nl/
Programing Logic	http://www.cl.cam.ac.uk/research/pls/
Security Group	http://www.cl.cam.ac.uk/research/security/
Systems Research	http://www.cl.cam.ac.uk/research/srg/
Distributed And Scientific Computing	http://www.cs.cf.ac.uk/research/dsc.php
Informatics	http://www.cs.cf.ac.uk/research/inf.php
Visual Computing	http://www.cs.cf.ac.uk/research/vis.php
Advanced Digital Manufacturing Technology	http://www.uclan.ac.uk/schools/computing/engineering/physical/admt/research.php
Digital Signal Processing	http://www.uclan.ac.uk/schools/computing/engineering/physical/adsip/research/index.php
Child Computer Interaction	http://www.uclan.ac.uk/schools/computing/engineering/physical/child/computer/interaction.php
Autonomous Intelligent Systems	http://www.city.ac.uk/informatics/school-organisation/department-of-computing/research/ais-group
Music Informatics	http://www.city.ac.uk/informatics/school-organisation/department-of-computing/research/music-informatics-group
Programming Languages And System	http://www.city.ac.uk/informatics/school-organisation/department-of-computing/research/plas-group
Software Engineering	http://www.city.ac.uk/informatics/school-organisation/department-of-computing/research/se-group
Pervasive Computing	http://wwwm.coventry.ac.uk/researchnet/cogentcomputing/Pages/CogentComputing.aspx
Distributed Systems	http://wwwm.coventry.ac.uk/researchnet/DistributedSystemsandModelling/Pages/DistributedSystemsandModelling.aspx
Interactive Words Research	http://wwwm.coventry.ac.uk/researchnet/iwarg/Pages/iWARG.aspx
Games And Virtual Worlds	http://wwwm.coventry.ac.uk/researchnet/sgarg/Pages/SeriousGamesandVirtualWorlds.aspx
Wireless Sensor Networks	http://wwwm.coventry.ac.uk/researchnet/cogentcomputing/Pages/CogentComputing.aspx
Digital Manufacturing	http://wwwm.coventry.ac.uk/researchnet/AdvancedDigitalManufacturing/Pages/AdvancedDigitalManufacturing.aspx
Electronics And Communication	http://www.dmu.ac.uk/research/research-faculties-and-institutes/technology/centre-for-electronic-and-comms-engineerineering/centre-for-electronic-and-communications-engineering.aspx
Cyber Security	http://www.dmu.ac.uk/research/research-faculties-and-institutes/technology/cyber-security-centre/cyber-security-centre.aspx
Computational Intelligence	http://146.227.2.132/
Imaging And Display	http://www.dmu.ac.uk/research-areas/idrg/research/context.jsp
Software	http://www.tech.dmu.ac.uk/STRL/research/areas/index.html

Technologies	
Assistive Health Care	http://www.computing.dundee.ac.uk/ac/research/groupdetails.asp?28
Computational Systems	http://www.computing.dundee.ac.uk/ac/research/groupdetails.asp?30
Interactive Systems Design	http://www.computing.dundee.ac.uk/ac/research/groupdetails.asp?27
Space Technology	http://www.computing.dundee.ac.uk/ac/research/groupdetails.asp?29
Algorithms And Complexity	http://www.dur.ac.uk/algorithms.complexity/
Innovative Computing	http://www.dur.ac.uk/ecs/ecs/research/researchstaff/innovativecomp/research/
Computational Biology	http://www.uea.ac.uk/cmp/research/cmpbio
Machine Learning	https://www.uea.ac.uk/cmp/research/mma/machineLearning
Statistics	https://www.uea.ac.uk/cmp/research/mma/statistics
Distributed Software Engineering	http://www.uel.ac.uk/ace/research/dse/
Information Security And Digital Forensics	http://www.uel.ac.uk/isdf/
Software Architecture	http://homepages.uel.ac.uk/R.Bashroush/SOAR/
Neuro Science	http://www.anc.ed.ac.uk/neuroscience
Machine Leaning	http://www.anc.ed.ac.uk/machine-learning
Bioinformatics	http://www.anc.ed.ac.uk/bioinformatics
Language Computation	http://www.ilcc.inf.ed.ac.uk/research
Machine Vision	http://www.ipab.inf.ed.ac.uk/mvu/
Robot Learning	http://wcms.inf.ed.ac.uk/ipab/slmc/welcome
Computer Graphics	http://www.ipab.inf.ed.ac.uk/cgvu/
Autonomy	http://wcms.inf.ed.ac.uk/ipab/autonomy/home
Intelligent Systems	http://www.cisa.inf.ed.ac.uk/
Compilers And Architecture	http://wcms.inf.ed.ac.uk/icsa/research/copy/of/icsa-research-groups
Parallel Computing	http://wcms.inf.ed.ac.uk/icsa/research/structured-parallelism-group
Wireless Networking	http://wcms.inf.ed.ac.uk/icsa/research/icsa-research-groups
Processor	http://groups.inf.ed.ac.uk/pasta/
Algorithms	http://www.lfcs.inf.ed.ac.uk/research/complexity/
Database	http://www.lfcs.inf.ed.ac.uk/research/database/
Security	http://www.lfcs.inf.ed.ac.uk/research/mobility+security/
Distributed Computing	http://www.cdcs.napier.ac.uk/
Emergent Computing	http://www.cec.napier.ac.uk/

Software Systems	http://www.ciss.napier.ac.uk/
Interactive Design	http://www.cid.napier.ac.uk/
Social Informatics	http://www.csi.napier.ac.uk/
Foundations And Application	http://www.essex.ac.uk/csee/research/groups/FoundApp/index.aspx
Future Networks	http://www.essex.ac.uk/csee/research/groups/FutureNet/index.aspx
High Performance Networks	http://hpn.essex.ac.uk/
Intelligent Systems	http://www.essex.ac.uk/csee/research/groups/IntSys/index.aspx
Multimedia	http://www.essex.ac.uk/csee/research/groups/Multimedia/index.aspx
Pervasive Systems	http://www.essex.ac.uk/csee/research/groups/Pervasive/index.aspx
Radio Frequencies	http://www.essex.ac.uk/csee/research/groups/RF/index.aspx
Robotics	http://www.essex.ac.uk/csee/research/groups/Robotics/index.aspx
Knowledge Representation	http://emps.exeter.ac.uk/mathematics-computer-science/research/computer-science/research-interests/knowledge-representation-ontology/
Machine Learning	http://emps.exeter.ac.uk/mathematics-computer-science/research/computer-science/research-interests/machine-learning/
Optimization	http://emps.exeter.ac.uk/mathematics-computer-science/research/computer-science/research-interests/optimisation/
Natural Computing	http://emps.exeter.ac.uk/mathematics-computer-science/research/computer-science/research-interests/nature-inspired-computing/
Shape Representation	http://emps.exeter.ac.uk/mathematics-computer-science/research/computer-science/research-interests/shape-representation/
Applied Mathematics	http://model.research.glam.ac.uk/
Data Integrity And Combinatorics	http://data.research.glam.ac.uk/
Game And AI	http://intelligence.research.glam.ac.uk/about/
Geographical Information Systems	http://gis.research.glam.ac.uk/about/
Hyper Media	http://hypermedia.research.glam.ac.uk/
Medical Imaging	http://imaging.research.glam.ac.uk/about/
Information Security	http://security.research.glam.ac.uk/
Medical And Signal Processing	http://mespru.research.glam.ac.uk/about/
Computer Vision	http://www.gla.ac.uk/schools/computing/research/researchgroups/computervisionandgraphics/
Embedded System	http://www.gla.ac.uk/schools/computing/research/researchgroups/embeddednetworkedanddistributedsystems/
Formal Analysis	http://www.gla.ac.uk/schools/computing/research/researchgroups/formalanalysistheoryandalgorithms/
Human Computer Interaction	http://www.gla.ac.uk/schools/computing/research/researchgroups/humancomputerinteractiongist/

Inference Dynamics	http://www.gla.ac.uk/schools/computing/research/researchgroups/inferencedynamicsandinteraction/
Information Retrieval	http://www.gla.ac.uk/schools/computing/research/researchgroups/informationretrieval/
Software Engineering	http://www.gla.ac.uk/schools/computing/research/researchgroups/softwareengineeringandinformationsecurity/
Interactive Communication	http://www.gcu.ac.uk/ebe/research/researchgroups/interactivecommunicationsandengineering/
Artificial Intelligence	http://cccs.gre.ac.uk/group/AI.html
Autonomics	http://cms1.gre.ac.uk/research/autonomics/
Biomedical	http://www.macs.hw.ac.uk/bisel/
Dependable Systems	http://www.macs.hw.ac.uk/~dsg/content/public/home/home.php
Intelligent Systems	http://www.macs.hw.ac.uk/isl/
Interaction Lab	http://sites.google.com/site/hwinteractionlab/
Pervasive And Mobile Computing	http://www.macs.hw.ac.uk/puma/
Texture Lab	http://www.macs.hw.ac.uk/texturelab/
Digital Media Processing	http://research-ecee.herts.ac.uk/
Optical Networks	http://www.herts.ac.uk/research-and-innovation/science-and-technology-research-institute/computer-science-and-informatics-research/onrg-optical-networks-research-group/home.cfm
Bio computation	http://homepages.feis.herts.ac.uk/~nngroup/
Adaptive Systems	http://adapsys.feis.herts.ac.uk/
Systems And Software	http://www.herts.ac.uk/research-and-innovation/science-and-technology-research-institute/computer-science-and-informatics-research/systems-and-software.cfm
Computer Graphics	http://www.hud.ac.uk/research/eps/cgiv/
Knowledge Engineering	http://www.hud.ac.uk/research/eps/keii/
Software Engineering	http://www.hud.ac.uk/research/eps/tserg/
Information Retrieval	http://www.hud.ac.uk/research/eps/xdir/
Distributed Systems	http://www2.hull.ac.uk/science/computer/science/research/dris.aspx
Simulation And Visualization	http://www2.hull.ac.uk/science/computer/science/research/simvis.aspx
Computer Architecture	http://comparch.doc.ic.ac.uk/
Large Scale Distributed Systems	http://lsds.doc.ic.ac.uk/research/
Security	http://www3.imperial.ac.uk/computing/research/security
Software	http://srg.doc.ic.ac.uk/
Computational Creativity	http://ccg.doc.ic.ac.uk/wiki/doku.php?id=main

Intelligent Behaviour Understanding	http://ibug.doc.ic.ac.uk/
Social Computing	http://scg.doc.ic.ac.uk/about/us
Bioinformatics	http://www.doc.ic.ac.uk/bioinformatics/
Computational Logic	http://www.doc.ic.ac.uk/~ft/argumentation.html
Machine Learning	http://www.doc.ic.ac.uk/research/machinelearning/DoC/Machine/Learning/ML.html
Neurodynamics	http://www.doc.ic.ac.uk/~mpsha/ComputationalNeurodynamicsGroup.html
Security	http://www3.imperial.ac.uk/computing/research/security
Autonomous System	http://vas.doc.ic.ac.uk/
Reasoning	http://www-lrr.doc.ic.ac.uk/
Reliable Web	http://www-rw.doc.ic.ac.uk/
Programming Languages	http://slurp.doc.ic.ac.uk/
Analysis Engineering	http://aesop.doc.ic.ac.uk/about/
Discovery Science	http://dsg.doc.ic.ac.uk/
Optimization	http://optimisation.doc.ic.ac.uk/
Embedded Systems	http://www.doc.ic.ac.uk/~es309/aese/
Software Engineering	http://www-dse.doc.ic.ac.uk/
Policy Based Autonomous Systems	http://www3.imperial.ac.uk/policyresearch
Biomedical Image	http://biomedic.doc.ic.ac.uk/
Medical Image Computing	http://ubimon.doc.ic.ac.uk/gzy/m375.html
Robot Vision	http://www2.imperial.ac.uk/robotvision/website/php/
Computational Intelligence	http://www.keele.ac.uk/scm/research/researchgroups/computationalintelligenceandcognitive/
Knowledge Modelling	http://www.keele.ac.uk/scm/research/researchgroups/knowledgemodelling/
Software Engineering	http://www.scm.keele.ac.uk/research/software/engineering/se/Research.html
Computational Intelligence	http://www.cs.kent.ac.uk/research/groups/compint/index.html
Computi8ng Education	http://www.cs.kent.ac.uk/research/groups/compedu/index.html
Future Computing	http://www.cs.kent.ac.uk/research/groups/future/index.html
Programming Languages	http://www.cs.kent.ac.uk/research/groups/plas/index.html
Security Group	http://www.cs.kent.ac.uk/research/groups/security/index.html
Bio imaging	http://sec.kingston.ac.uk/research/research-groups/big/
Bioinformatics	http://sec.kingston.ac.uk/research/research-groups/biogsp/

Distributed Systems	http://sec.kingston.ac.uk/research/research-groups/codis/
Digital Imaging	http://sec.kingston.ac.uk/research/research-centres/digital-imaging-research-centre/
Emerging Technologies	http://business.kingston.ac.uk/research/research-groups/emerging-technologies-group
Human Body Motion	http://sec.kingston.ac.uk/research/research-groups/hbm/
Learning Technology	http://sec.kingston.ac.uk/research/research-groups/ltg/
Mobile Information	http://sec.kingston.ac.uk/research/research-centres/mobile-information-and-network-technologies/
Mobile Information For Embedded Systems	http://sec.kingston.ac.uk/research/research-groups/momed/
Robot Vision	http://sec.kingston.ac.uk/research/research-groups/rovit/
Scientific Analysis And Visualization	http://sec.kingston.ac.uk/research/research-groups/savic/
User Experience	http://sec.kingston.ac.uk/research/research-groups/user-experience/
Visual Surveillance	http://sec.kingston.ac.uk/research/research-groups/vsrg/
Wireless Multimedia Networking	http://sec.kingston.ac.uk/research/research-groups/wmn/
Research	http://www.scc.lancs.ac.uk/research/
Computer Corpus Research	http://ucrel.lancs.ac.uk/
Bio systems	http://www.comp.leeds.ac.uk/biosystems/
Computer Vision	http://www.comp.leeds.ac.uk/vision/
Knowledge Representation	http://www.comp.leeds.ac.uk/krr/
Medical Image	http://www.comp.leeds.ac.uk/vision/med/image.html
Natural Language Processing	http://www.comp.leeds.ac.uk/nlp/
Algorithms	http://www.engineering.leeds.ac.uk/computing/research/algorithms/
Distributed Systems	http://www.comp.leeds.ac.uk/distsys/
Collaborative Systems	http://www.comp.leeds.ac.uk/CollabSysAndPerf/
Scientific Computation	http://www.comp.leeds.ac.uk/scicomp/
Visualization And Reality	http://www.comp.leeds.ac.uk/vvr/
Research Themes	http://www2.le.ac.uk/departments/computer-science/research/themes
Vision And Robotics	http://www.lincoln.ac.uk/socs/research/vissur/default.htm
Robotics	http://robots.lincoln.ac.uk/index.html
Digital Contents	http://dcapi.blogs.lincoln.ac.uk/
Agents	http://www.csc.liv.ac.uk/research/agents/
Algorithms	http://www.csc.liv.ac.uk/research/ctag/
Logic	http://www.csc.liv.ac.uk/research/logics/

Economics And Computation	http://www.csc.liv.ac.uk/research/ecco/
Intelligent Distributed Systems	http://www.hope.ac.uk/intelligent-and-distributed-systems-laboratory/intelligent-and-distributed-systems-laboratory.html
Intelligent Systems	http://www.dcs.bbk.ac.uk/research/compint/
Web Technologies	http://www.dcs.bbk.ac.uk/research/dbtech/
Knowledge Lab	http://www.lkl.ac.uk/cms/index.php?option=com/content&task=blogcategory&id=41&Itemid=109
Agents	http://www.kcl.ac.uk/nms/depts/informatics/research/ais/index.aspx
Bio Informatics	http://www.kcl.ac.uk/nms/depts/informatics/research/bad/index.aspx
Robotics	http://www.kcl.ac.uk/nms/depts/engineering/research/Robotics/index.aspx
Telecommunications	http://www.kcl.ac.uk/nms/depts/engineering/research/telecommunications/index.aspx
Wide Band Communications	http://www.kcl.ac.uk/nms/depts/engineering/research/ultrawidebandcommunications/index.aspx
Theoretical Computer Science	http://www.dcs.qmul.ac.uk/research/logic/QM-EECS-TCS/Welcome.html
Digital Music	http://www.elec.qmul.ac.uk/digitalmusic/index.html
Multimedia	http://www.elec.qmul.ac.uk/mmv/
Networks Communication	http://www.lboro.ac.uk/departments/co/research/nccs.html
Vision Imaging	http://www.lboro.ac.uk/departments/co/research/vias.html
Intelligent Interactive Systems	http://www.lboro.ac.uk/departments/co/research/iis.html
Theoretical Computer Science	http://www.lboro.ac.uk/departments/co/research/tcs.html
Advanced Processor	http://apt.cs.manchester.ac.uk/
Bio Health Informatics	http://bhig.cs.manchester.ac.uk/
Formal Methods	http://www.cs.manchester.ac.uk/research/groups/foundations/
Information Management	http://img.cs.manchester.ac.uk/
Machine Learning And Optimization	http://mlo.cs.man.ac.uk/
Nano Engineering	http://nest.cs.manchester.ac.uk/
Software Systems	http://www.cs.manchester.ac.uk/research/groups/softwaresystems/
Text Mining	http://www.cs.manchester.ac.uk/research/groups/infosystems/textmining/
Advance Interfaces	http://aig.cs.man.ac.uk/home/home.php
Imaging	http://www.medicine.manchester.ac.uk/imaging/
Mathematical Modelling	http://www.scmdt.mmu.ac.uk/cmmfa/
Distributed Networks	http://www.scmdt.mmu.ac.uk/research/funds/

Image Sensoring	http://www.scmdt.mmu.ac.uk/RESEARCH/ISCA/
Intelligent Systems	http://www.scmdt.mmu.ac.uk/RESEARCH/Intelgrp/
Computational Logic	http://www.scmdt.mmu.ac.uk/RESEARCH/logicgrp/
Novel Computation Group	http://www.scmdt.mmu.ac.uk/RESEARCH/ncg/index.html
Automated Scheduling	http://www.asap.cs.nott.ac.uk/
Algorithmic Problem Solving	http://aps.cs.nott.ac.uk/
Functional Programming	http://sneezy.cs.nott.ac.uk/joomla/
Agents Lab	http://www.agents.cs.nott.ac.uk/
Intelligent Modelling	http://ima.ac.uk/
Computing And Complex Systems	http://icos.cs.nott.ac.uk/
Mixed Reality	http://www.mrl.nott.ac.uk/mrl-research.html
Applied Image And Display	http://www.ntu.ac.uk/research/groups/centres/sat/81247.html
Communication Systems	http://www.ntu.ac.uk/research/groups/centres/sat/82250.html
Computational Optimisation	http://www.ntu.ac.uk/research/groups/centres/sat/91866.html
Intelligent Recognition	http://www.ntu.ac.uk/research/groups/centres/sat/81243.html
Intelligent Simulation	http://www.ntu.ac.uk/research/groups/centres/sat/81244.html
Interactive Systems	http://www.ntu.ac.uk/research/groups/centres/sat/81245.html
Computational Biology	http://www.cs.ox.ac.uk/research/compbio/
Foundations And Logic	http://www.cs.ox.ac.uk/research/fls/
Information Systems	http://www.cs.ox.ac.uk/research/is/
Programming Languages	http://www.cs.ox.ac.uk/research/pl/
Software Engineering	http://www.cs.ox.ac.uk/research/se/
Verification	http://www.cs.ox.ac.uk/research/verification/
Brookes Research	http://cct.brookes.ac.uk/research/computer-science/index.html
Robotics And Neural Systems	http://www.plymouth.ac.uk/research/crns
Security Communications And Networks	http://www.plymouth.ac.uk/research/cscan
Cognitive Systems Engineering	http://www.port.ac.uk/research/serg/cognitivesystemsengineering/
Healthcare Modelling	http://www.chmi.port.ac.uk/

And Informatics	
Systems Engineering	http://www.port.ac.uk/research/serg/
Systems Information	http://www.port.ac.uk/departments/academic/comp/research/sis/
Advanced Networks	http://www.ecit.qub.ac.uk/Research/DigitalCommunications/AdvancedNetworks/
Digital Signal Processing	http://www.ecit.qub.ac.uk/Research/DigitalCommunications/DSPandCommunications/
Programmable Systems And Networks	http://www.ecit.qub.ac.uk/Research/DigitalCommunications/ProgrammableSystemsandNetworks/
Radio Communications	http://www.ecit.qub.ac.uk/Research/DigitalCommunications/RadioCommunication/
Wireless Networking	http://www.ecit.qub.ac.uk/Research/DigitalCommunications/WirelessNetworking/
Distributed Computing	http://www.qub.ac.uk/research-centres/HPDC/
Intelligent Systems	http://www.qub.ac.uk/research-centres/ISAC/Introduction/
Data Engineering	http://www.qub.ac.uk/research-centres/KDE/
Computing	http://www.reading.ac.uk/sse/research/sse-computing.aspx
Cybernetics	http://www.reading.ac.uk/sse/research/sse-cybernetics.aspx
Communications	http://www.reading.ac.uk/sse/research/sse-information-and-communications.aspx
Intelligent Systems	http://www.isr.reading.ac.uk/
Computational Algebra	http://www.cs.st-andrews.ac.uk/research/aisc/cacl
Constraint Programming	http://www.cs.st-andrews.ac.uk/research/aisc/cp
Cognitive Systems	http://www.cs.st-andrews.ac.uk/research/aisc/cs
Distributed Systems	http://www.cs.st-andrews.ac.uk/research/nds
Systems Engineering	http://www.cs.st-andrews.ac.uk/research/se
Networking And Communications	http://www.cse.salford.ac.uk/research/networking-telecommunications/
Data Mining And Pattern Recognition	http://www.cse.salford.ac.uk/research/data-mining-pattern-recognition/
Autonomous Systems	http://www.cse.salford.ac.uk/research/autonomous-systems-robotics/themes.php
Comp Biology	http://www.sheffield.ac.uk/dcs/research/groups/compbio
Comp Graphics	http://www.sheffield.ac.uk/dcs/research/groups/graphics
Machine Learning	http://ml.dcs.shef.ac.uk/
Natural Language Processing	http://nlp.shef.ac.uk/
Neuro Computing	http://www.sheffield.ac.uk/dcs/research/groups/nrg
Organization Of Information Knowledge	http://oak.dcs.shef.ac.uk/?q=research

Speech And Hearing	http://spandh.dcs.shef.ac.uk/
Verification And Testing	http://www.sheffield.ac.uk/dcs/research/groups/vt
Agents Interaction	http://www.aic.ecs.soton.ac.uk/
Communication And Signal Processing	http://www.cspc.ecs.soton.ac.uk/research/areas
Software Systems	http://www.ess.ecs.soton.ac.uk/about
Web Science	http://www.wais.ecs.soton.ac.uk/researchthemes
Security Systems	http://www.staffs.ac.uk/faculties/comp/eng/tech/research/ciiss/
Mobile Fusion	http://www.staffs.ac.uk/faculties/comp/eng/tech/research/mfc/
Planning	http://planning.cis.strath.ac.uk/
Software Systems	http://ssg.cis.strath.ac.uk/
Structured Programming	http://www.msp.cis.strath.ac.uk/
Mobile Group	http://web.me.com/richard.c.connor/GMDS/GMDS/group.html
Mobiquitous	http://mobiquitous.cis.strath.ac.uk/
Combinatorics	http://combinatorics.cis.strath.ac.uk/research/
Formal Methods And Security	http://www.surrey.ac.uk/computing/research/fms/
Digital Ecosystems	http://www.surrey.ac.uk/computing/research/de/
Nature Inspired Computing	http://www.surrey.ac.uk/computing/research/nice/
Multimedia Security And Forensics	http://www.surrey.ac.uk/computing/research/msf/
Cognitive And Language Processing	http://www.sussex.ac.uk/calps/research
Evolutionary And Adaptive Systems	http://www.sussex.ac.uk/informatics/research/evolutionaryandadaptive systems
Foundations Of Software Systems	http://www.informatics.sussex.ac.uk/research/groups/FoSS/research/
Human Centered Technology	http://www.informatics.sussex.ac.uk/research/groups/hct/
Music Informatics	http://www.sussex.ac.uk/Users/nc81/research.html
Center For Computer Graphics	http://www.informatics.sussex.ac.uk/research/groups/cvcg/
Metropolitan Communications And Networking	http://www.smu.ac.uk/research/index.php/communications-and-networking
Metropolitan Pedagogy	http://www.smu.ac.uk/research/index.php/pedagogy-in-computing
Metropolitan Medical Signal Processing	http://www.smu.ac.uk/research/index.php/medical-signal-processing
Metropolitan Computer Games	http://www.smu.ac.uk/research/index.php/computer-games
Metropolitan Prism	http://www.smu.ac.uk/research/index.php/prism

Artificial Intelligence	http://www.compeng.ulster.ac.uk/ai/rg.php
Information And Communication	http://www.compeng.ulster.ac.uk/ise/rg.php
Intelligent Systems	http://www.compeng.ulster.ac.uk/is/rg.php
Advanced Computer Architecture	http://www.cs.york.ac.uk/research/research-groups/aca/
Enterprise Systems	http://www.cs.york.ac.uk/research/research-groups/es/
Non Standard Computation	http://www.cs.york.ac.uk/research/research-groups/nsc/
Artificial Intelligence	http://www.cs.york.ac.uk/research/research-groups/ai/
Real Time Systems	http://www.cs.york.ac.uk/research/research-groups/rts/
Computer Vision	http://www.cs.york.ac.uk/research/research-groups/cvpr/
High Integrity Systems	http://www.cs.york.ac.uk/research/research-groups/hise/
Programming Languages	http://www.cs.york.ac.uk/research/research-groups/plasma/
Human Computer Interaction	http://www.cs.york.ac.uk/research/research-groups/hci/
Research	http://www.wlv.ac.uk/default.aspx?page=26015
Audio Visual Communications	http://www.uws.ac.uk/schoolsdepts/computing/Research/avcn/areas.asp
Database And Semantic	http://www.uws.ac.uk/schoolsdepts/computing/Research/database-research-group/index.asp
ICT In Education	http://icte.uws.ac.uk/
Virtual Worlds	http://www.uws.ac.uk/schoolsdepts/computing/Research/VWVLEProject.asp
Applied DSP	http://www.westminster.ac.uk/research/a-z/applied-dsp-and-vlsi
Parallel Computing	http://www.westminster.ac.uk/research/a-z/centre-for-parallel-computing
System Analysis	http://www.westminster.ac.uk/research/a-z/centre-for-systems-analysis
Communications And Compunetics	http://www.westminster.ac.uk/research/a-z/communications-and-compunetics
Computational Intelligence	http://www.westminster.ac.uk/research/a-z/computational-intelligence-group
Data Knowledge Management	http://www.westminster.ac.uk/research/a-z/data-and-knowledge-management
Distributed Intelligent Systems	http://www.westminster.ac.uk/research/a-z/distributed-and-intelligent-systems
Health And Social Care Modelling	http://www.westminster.ac.uk/research/a-z/health-and-social-care-modelling
Mobile And Wireless Computing	http://www.westminster.ac.uk/research/a-z/mobile-and-wireless-computing
Software Systems Engineering	http://www.westminster.ac.uk/research/a-z/software-systems-engineering
Systems	http://www.westminster.ac.uk/research/a-z/systems-interoperability

Interoperability	
Wireless Communications	http://www.westminster.ac.uk/research/a-z/wireless-communications
Computational Biology	http://www2.warwick.ac.uk/fac/sci/dcs/research/combi/
Foundations Of Computer Science	http://www2.warwick.ac.uk/fac/sci/dcs/research/focs/
Intelligent And Adaptive Systems	http://www2.warwick.ac.uk/fac/sci/dcs/research/ias
Performance And Computing Visualization	http://www2.warwick.ac.uk/fac/sci/dcs/research/pcav/
Discrete Mathematics	http://www2.warwick.ac.uk/fac/cross/fac/dimap

Appendix C: Geo location of the 36 UK Universities in 2013 Leiden Ranking

Website	**Institution**	**Address**	**Latitude**	**Longitude**
shef.ac.uk	The University of Sheffield	The University of Sheffield, Western Bank, Sheffield, South Yorkshire S10 2TN, UK	53.3809409	-1.4879469
st-andrews.ac.uk	University of St Andrews	University of St Andrews, Saint Andrews, Fife KY16 9AJ, UK	56.3416934	-2.7927522
gla.ac.uk	University of Glasgow	University of Glasgow, Glasgow, Glasgow City G12 8QQ, UK	55.8721211	-4.2882005
le.ac.uk	University of Leicester	University of Leicester, University Road, Leicester LE1 7RH, UK	52.6211393	-1.1246325
uea.ac.uk	University of East Anglia	University of East Anglia, Norwich Research Park, Norwich, Norfolk NR4 7TJ, UK	52.6219215	1.2391761
ox.ac.uk	University of Oxford	University of Oxford, University Offices, Wellington Square, Oxford	51.7566341	-1.2547037

		OX1 2JD, UK		
rdg.ac.uk	University of Reading	University of Reading, Reading, Berkshire, UK	51.4414205	-0.9418157
surrey.ac.uk	University of Surrey	University of Surrey, Guildford, Surrey GU2 7XH, UK	51.242722	-0.5895144
lshtm.ac.uk	London School of Hygiene and Tropical Medicine	London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK	51.5206139	-0.1300205
leeds.ac.uk	University of Leeds	University of Leeds, Leeds, West Yorkshire LS2 9JT, UK	53.8066815	-1.5550328
ucl.ac.uk	University College London	University College London, Gower Street, London WC1E 6BT, UK	51.5245592	-0.1340401
ex.ac.uk	University of Exeter	University of Exeter, Exeter, Devon EX4, UK	50.7371369	-3.5351475
imperial.ac.uk	Imperial College	Imperial College, Kensington, London SW7 2AZ, UK	51.4987835	-0.1748876
warwick.ac.uk	University of Warwick	University of Warwick, Coventry CV4 7AL, UK	52.3792525	-1.5614704
cardiff.ac.uk	Cardiff University	Cardiff University, Cardiff CF10 3XQ, UK	51.4866271	-3.1788641
nott.ac.uk	The University Of Nottingham - School of Nursing	The University Of Nottingham - School of Nursing, Mansfield Road, Sutton-in-Ashfield, Nottinghamshire NG17 4JL, UK	53.134511	-1.236258
susx.ac.uk	University of Sussex	University of Sussex, Brighton BN1 9RH, UK	50.8670895	-0.087914
kcl.ac.uk	King's College London	King's College London, 20 Newcomen Street, London SE1 1UL, UK	51.5033351	-0.0897397
qmul.ac.uk	Queen Mary University of London	Queen Mary University of London, Mile End Rd, London E1	51.5240671	-0.0403745

		4NS, UK		
lancs.ac.uk	Lancaster University	Lancaster University, Bailrigg, Lancaster, Lancashire LA1 4YW, UK	54.0103942	-2.7877294
qub.ac.uk	Queens University	Queens University, Belfast BT9 5NB, UK	54.5537538	-5.9666672
man.ac.uk	The University of Manchester	The University of Manchester, Oxford Road, Manchester M13 9PL, UK	53.4665323	-2.2335496
dur.ac.uk	Durham University	Durham University, Stockton Rd, Durham, County Durham DH1, UK	54.7649859	-1.5782029
strath.ac.uk	University of Strathclyde	University of Strathclyde, 16 Richmond Street, Glasgow, Glasgow City G1 1XQ, UK	55.8624195	-4.2425876
dundee.ac.uk	University of Dundee	University of Dundee, Nethergate, Dundee, Dundee City DD1 4HN, UK	56.4582447	-2.9821428
abdn.ac.uk	University of Aberdeen	University of Aberdeen, University of Aberdeen King's Campus, King's College, Aberdeen, Aberdeen City AB24 3FX	57.1650429	-2.1002589
york.ac.uk	University of York	University of York, Heslington, York YO10 5DD, UK	53.9455334	-1.0561667
ncl.ac.uk	Newcastle University	Newcastle University, Newcastle upon Tyne, Tyne and Wear NE1 7RU, UK	54.9794793	-1.6147435
ed.ac.uk	The University of Edinburgh	The University of Edinburgh, Old College, South Bridge, Edinburgh, City of Edinburgh EH8 9YL, UK	55.9445158	-3.1892413
cam.ac.uk	University of Cambridge	University of Cambridge, The Old Schools, Trinity Lane, Cambridge, Cambridgeshire	52.2042666	0.1149085

		CB2 1TN, UK		
soton.ac.uk	University of Southampton Highfield Campus	University of Southampton Highfield Campus, Southampton SO17 1BJ	50.935742	-1.3966381
bris.ac.uk	University of Bristol	University of Bristol, Senate House, Tyndall Avenue, Bristol, City of Bristol BS8 1TH, UK	51.4584172	-2.6029792
lboro.ac.uk	Loughborough University	Loughborough University, Loughborough, Leicestershire LE11 3TU, UK	52.7641408	-1.2333126
bath.ac.uk	University of Bath	University of Bath, Bath, North East Somerset BA2 7AY, UK	51.3777431	-2.3263779
bham.ac.uk	University of Birmingham	University of Birmingham, Birmingham, West Midlands B15 2TT, UK	52.4508168	-1.9305135
liv.ac.uk	University of Liverpool	University of Liverpool, Liverpool, Merseyside L69 3BX, UK	53.405936	-2.9655722

Appendix D: An example of the computation of geographic distance

The geographic distance separating gre(55.8721211 -4.2882005) and shef(53.3809409 -1.4879469).

gla.ac.uk to radians = (0.975152473 -0.0748432177)

shef.ac.uk to radians = (0.931673177 -0.0259695725)

Radius of earth = 6371km

$$distance\ between\ gla\ and\ shef = \Big(\arccos\Big(\sin(0.975152473) * \sin(0.931673177) + \cos(0.975152473) * \cos(0.931673177) * cos\big(-0.0259695725 - (-0.0748432177)\big)\Big)\Big) * 6371 = 330.43k\text{m}$$

Appendix E: Search engine queries for the number times the University of York is mentioned in other UK university websites.

"university of york" site:abdn.ac.uk	"york university" site:abdn.ac.uk
"university of york" site:bath.ac.uk	"york university" site:bath.ac.uk
"university of york" site:birmingham.ac.uk	"york university" site:birmingham.ac.uk
"university of york" site:bristol.ac.uk	"york university" site:bristol.ac.uk
"university of york" site:cam.ac.uk	"york university" site:cam.ac.uk
"university of york" site:cardiff.ac.uk	"york university" site:cardiff.ac.uk
"university of york" site:dundee.ac.uk	"york university" site:dundee.ac.uk
"university of york" site:dur.ac.uk	"york university" site:dur.ac.uk

"university of york" site:ed.ac.uk	"york university" site:ed.ac.uk
"university of york" site:exeter.ac.uk	"york university" site:exeter.ac.uk
"university of york" site:gla.ac.uk	"york university" site:gla.ac.uk
"university of york" site:imperial.ac.uk	"york university" site:imperial.ac.uk
"university of york" site:kcl.ac.uk	"york university" site:kcl.ac.uk
"university of york" site:lancaster.ac.uk	"york university" site:lancaster.ac.uk
"university of york" site:lboro.ac.uk	"york university" site:lboro.ac.uk
"university of york" site:le.ac.uk	"york university" site:le.ac.uk
"university of york" site:leeds.ac.uk	"york university" site:leeds.ac.uk
"university of york" site:liv.ac.uk	"york university" site:liv.ac.uk
"university of york" site:lshtm.ac.uk	"york university" site:lshtm.ac.uk
"university of york" site:manchester.ac.uk	"york university" site:manchester.ac.uk
"university of york" site:ncl.ac.uk	"york university" site:ncl.ac.uk
"university of york" site:nottingham.ac.uk	"york university" site:nottingham.ac.uk
"university of york" site:ox.ac.uk	"york university" site:ox.ac.uk
"university of york" site:qmul.ac.uk	"york university" site:qmul.ac.uk
"university of york" site:qub.ac.uk	"york university" site:qub.ac.uk
"university of york" site:reading.ac.uk	"york university" site:reading.ac.uk
"university of york" site:sheffield.ac.uk	"york university" site:sheffield.ac.uk
"university of york" site:southampton.ac.uk	"york university" site:southampton.ac.uk
"university of york" site:st-andrews.ac.uk	"york university" site:st-andrews.ac.uk
"university of york" site:strath.ac.uk	"york university" site:strath.ac.uk
"university of york" site:surrey.ac.uk	"york university" site:surrey.ac.uk
"university of york" site:sussex.ac.uk	"york university" site:sussex.ac.uk
"university of york" site:ucl.ac.uk	"york university" site:ucl.ac.uk
"university of york" site:uea.ac.uk	"york university" site:uea.ac.uk
"university of york" site:warwick.ac.uk	"york university" site:warwick.ac.uk

**********THE END**********